## econometrics

## Celebrated

# Econometricians 

## Katarina Juselius and

 Soren JohansenEdited by
Rocco Mosconi and Paolo Paruolo
Printed Edition of the Special Issue Published in Econometrics

# Celebrated Econometricians: Katarina Juselius and Søren Johansen 

# Celebrated Econometricians: Katarina Juselius and Søren Johansen 

Editors

Rocco Mosconi
Paolo Paruolo

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin

Editors

Rocco Mosconi
Politecnico di Milano Italy

Paolo Paruolo
European Commission, Joint
Research Centre
Italy

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal Econometrics (ISSN 2225-1146) (available at: https://www.mdpi.com/journal/econometrics/special_ issues/Juselius_Johansen).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. Journal Name Year, Volume Number, Page Range.

ISBN 978-3-0365-4969-9 (Hbk)
ISBN 978-3-0365-4970-5 (PDF)

Cover image courtesy of https:/ / pxhere.com/en/photo/99262
© 2022 by the authors. Articles in this book are Open Access and distributed under the Creative Commons Attribution (CC BY) license, which allows users to download, copy and build upon published articles, as long as the author and publisher are properly credited, which ensures maximum dissemination and a wider impact of our publications.
The book as a whole is distributed by MDPI under the terms and conditions of the Creative Commons license CC BY-NC-ND.

## Contents

About the Editors ..... vii
Rocco Mosconi, Paolo Paruolo
Celebrated Econometricians: Katarina Juselius and Søren Johansen
Reprinted from: Econometrics 2022, 10, 24, doi:10.3390/econometrics10020024 ..... 1
Rocco Mosconi, Paolo Paruolo
A Conversation with Katarina Juselius
Reprinted from: Econometrics 2022, 10, 20, doi:10.3390/econometrics10020020 ..... 5
Rocco Mosconi, Paolo Paruolo
A Conversation with Søren Johansen
Reprinted from: Econometrics 2022, 10, 21, doi:10.3390/econometrics10020021 ..... 27
Katarina JuseliusSearching for a Theory That Fits the Data: A Personal Research OdysseyReprinted from: Econometrics 2021, 9, 5, doi:10.3390/econometrics901000543
Fragiskos Archontakis, Rocco Mosconi
Søren Johansen and Katarina Juselius:A Bibliometric Analysis of Citations through Multivariate Bass Models
Reprinted from: Econometrics 2021, 9, 30, doi:10.3390/econometrics9030030 ..... 71
Søren Johansen
Cointegration and Adjustment in the $\operatorname{CVAR}(\infty)$ Representation of SomePartially Observed CVAR(1) Models
Reprinted from: Econometrics 2019, 7, 2, doi:10.3390/econometrics7010002 ..... 99
Matteo Barigozzi, Marco Lippi and Matteo Luciani
Cointegration and Error Correction Mechanisms for Singular Stochastic Vectors Reprinted from: Econometrics 2020, 8, 3, doi:10.3390/econometrics8010003 ..... 109
Dietmar Bauer, Lukas Matuschek, Patrick de Matos Ribeiro and Martin Wagner
A Parameterization of Models for Unit Root Processes: Structure Theory and Hypothesis Testing
Reprinted from: Econometrics 2020, 8, 42, doi:10.3390/econometrics8040042 ..... 133
Massimo Franchi and Paolo Paruolo
Cointegration, Root Functions and Minimal Bases
Reprinted from: Econometrics 2021, 9, 31, doi:10.3390/econometrics9030031 ..... 187
Bruce E. Hansen
Johansen's Reduced Rank Estimator Is GMM
Reprinted from: Econometrics 2018, 6, 26, doi:10.3390/econometrics6020026 ..... 215
David H. Bernstein and Bent Nielsen
Asymptotic Theory for Cointegration Analysis When the Cointegration Rank Is Deficient Reprinted from: Econometrics 2019, 7, 6, doi:10.3390/econometrics7010006 ..... 225
Takamitsu Kurita and Bent Nielsen
Partial Cointegrated Vector Autoregressive Models with Structural Breaks in Deterministic Terms Reprinted from: Econometrics 2019, 7, 42, doi:10.3390/econometrics8040042 ..... 249
Yuanyuan Li and Dietmar Bauer
Modeling I(2) Processes Using Vector Autoregressions Where the Lag Length Increases with the Sample Size
Reprinted from: Econometrics 2020, 8, 38, doi:10.3390/econometrics8030038 ..... 285
Andreas Hetland
The Stochastic Stationary Root Model
Reprinted from: Econometrics 2018, 6, 39, doi:10.3390/econometrics6030039 ..... 313
Niels Haldrup and and Carsten P. T. Rosenskjold
A Parametric Factor Model of the Term Structure of Mortality
Reprinted from: Econometrics 2019, 7, 9, doi:10.3390/econometrics7010009 ..... 347
Kevin D. Hoover
The Discovery of Long-Run Causal Order: A Preliminary Investigation Reprinted from: Econometrics 2020, 8, 31, doi:10.3390/econometrics 9030031 ..... 369
Jennifer L. Castle, David F. Hendry and Andrew B. Martinez
Evaluating Forecasts, Narratives and Policy Using a Test of Invariance
Reprinted from: Econometrics 2017, 5, 39, doi:10.3390/econometrics5030039 ..... 393
Helmut Lütkepohl and Aleksei Netšunajev
The Relation between Monetary Policy and the Stock Market inEurope Reprinted from: Econometrics 2018, 6, 36, doi:10.3390/econometrics6030036 ..... 421
Marit Gjelsvik, Ragnar Nymoen and Victoria Sparrman
Cointegration and Structure in Norwegian Wage-Price Dynamics ${ }^{\dagger}$
Reprinted from: Econometrics 2020, 8, 29, doi:10.3390/econometrics8030029 ..... 435
Michael D. Goldberg, Olesia Kozlova and Deniz Ozabaci
Forward Rate Bias in Developed and Developing Countries: More RiskyNot Less Rational Reprinted from: Econometrics 2020, 8, 43, doi:10.3390/econometrics8040043 ..... 451

## About the Editors

## Rocco Mosconi

Rocco Mosconi, MSc in Economics, University of Pavia. Currently full Professor of Econometrics, Department of Management, Economics and Industrial Engineering, Politecnico di Milano (since March 2006). His research interests include theoretical econometrics (time series analysis, non-stationary time series modeling, VAR models, survival data analysis, discrete choice models) and applied econometrics (regarding finance, industrial economics, and energy). He has published in the Journal of Econometrics, Econometrics Journal, Oxford Bulletin of Economics and Statistics, Econometrics, The Journal of Industrial Economics, Strategic Management Journal, Energy Policy.

## Paolo Paruolo

Paolo Paruolo has a PhD in Mathematical Statistics (University of Copenhagen Paruolo 1995). He has worked as a Professor of Econometrics (University of Bologna until 1999 and University of Insubria, 1999-2013). Currently, he works as a Scientific officer at the Joint Research Centre, European Commission, 2013-). He was ranked among the best 150 econometricians worldwide in Baltagi (2007) "Worldwide econometrics rankings: 1989-2005", Econometric Theory 23, p. 952-1012. He has published in: Econometrica, Journal of Econometrics, Econometric Reviews, Computational Economics, Economic Modelling, Journal of Economic Behavior \& Organization, Journal of the Royal Statistical Society A, Linear Algebra and its Applications, Journal of Geophysical Research-Atmospheres.

## Editorial

# Celebrated Econometricians: Katarina Juselius and Søren Johansen 

Rocco Mosconi ${ }^{1}$ and Paolo Paruolo ${ }^{2, *}$<br>1 Dipartimento di Ingegneria Gestionale, Politecnico di Milano, Via Lambruschini 4/B, 20156 Milano, Italy; rocco.mosconi@polimi.it<br>2 European Commission, Joint Research Centre (JRC), Via E. Fermi 2749, 21027 Ispra, Italy<br>* Correspondence: paolo.paruolo@ec.europa.eu

Citation: Mosconi, Rocco, and Paolo Paruolo. 2022. Celebrated Econometricians: Katarina Juselius and Søren Johansen. Econometrics 10: 24. https://doi.org/10.3390/ econometrics10020024

Received: 5 May 2022
Accepted: 9 May 2022
Published: 16 May 2022
Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/).

This Special Issue collects contributions related to the advances in the theory and practice of Econometrics induced by the research of Katarina Juselius and Søren Johansen, whom this Special Issue aims to celebrate.

The research of Katarina and Søren has been advancing Econometrics on fundamental issues, such as on common trends, equilibrium relations, adjustment to the (dis-)equilibrium relations, rationality of agents and on the discussion of resulting policy recommendations. Their research addressed issues of representation, identification, estimation, inference and policy implications, developing methodology and providing inspiring and paradigmatic applications in several applied areas of Economics.

One main body of work in Katarina's and Søren's research concerns Cointegration analysis using Vector Autoregressions (VAR), often referred to CVARs, both when the variables are integrated of order $1(\mathrm{I}(1))$ and $2(\mathrm{I}(2))$. Their contributions go beyond CVARs and have a very wide range, which is also partly reflected in the contributions of this Special Issue.

As a collection, the papers appearing in this Special Issue continue this tradition by providing advances on several topics, many of them related to the econometric analysis of nonstationary time-series. At the same time, from a complementary angle they also offer a recent perspective on the scope, breath and importance of some of the contributions of Katarina and Søren to Econometrics.

The papers in this Special Issue are both theoretical and applied, and they are grouped in the following areas for simplicity of exposition in this editorial. A first group of papers provides a historical perspective on Katarina's and Søren's contributions to Econometrics. A second group concentrates attention on representation theory; a third one focuses on estimation and inference. A fourth one deals with extensions of CVARs for modeling and forecasting, and a final fifth group is centered on empirical applications. These groups of papers are reviewed below; a final section of this editorial is dedicated to our many thanks associated with the preparation of this Special Issue.

## 1. A Historical Perspective

A first set of four papers, Archontakis and Mosconi (2021), Juselius (2021), Mosconi and Paruolo (2022a, 2022b), focuses on some of the contributions from Katarina and Søren to Econometrics, especially on early developments of cointegration.

Two separate interviews (Mosconi and Paruolo 2022a, 2022b), offer the reader a glimpse of Katarina's and Søren's motivation, hurdles and accomplishments in developing their research agenda. While several other joint interviews of Katarina and Søren exist, the ones in this Special Issue focus on their distinct contributions and hopefully provide a better account of their personal points of view.

Katarina's paper (Juselius 2021) is a complement to her interview in Mosconi and Paruolo (2022a); in this paper she gives account of her 'Research Odyssey' associated with the idea to understand macroeconomic data. She discusses rational and imperfect
knowledge expectations and how to learn from the many periods of crisis. The paper gives a concise but comprehensive overview of Katarina's model building approach based on "searching for a theory that fits the data" rather than "data that fits the theory".

Archontakis and Mosconi (2021) provide an bibliometric analysis on Katarina's and Søren's publications using a multivariate Bass model. They distinguish methodological and applied papers citing Katarina's and Søren's research and find cross-fertilization between the two areas. They show that the number of applied papers per quarter citing Katarina's and Søren's work does not seem to have peaked yet, while the methodological literature referring to their work reached the peak after the turn of the century, with a flat trajectory after the maximum (a similar behavior is observed in a minority of Nobel prize winners, and it is defined as "staying power" in the literature).

## 2. Representation

A second set of four papers (Barigozzi et al. 2020; Bauer et al. 2020; Franchi and Paruolo 2021; Johansen 2019) is concerned with representation theory, which plays a central role in Cointegration. An example of this is Granger's Representation Theorem, which shows that Cointegration (Common trends) and Equilibrium Corrections Mechanism (ECM) are dual concepts.

Søren's paper (Johansen 2019) derives the CVAR $(\infty)$ representation, and the corresponding finite order approximation, for a subset of observed variables generated by a higher dimensional CVAR model with lag order 1, which also includes a set of unobserved strongly exogenous random walks. The paper discusses cointegration, non-causality and weak exogeneity conditions for the observed variables and is motivated by some of the hypotheses proposed in Hoover (2020) in this Special Issue. The two papers allow to connect more explicitly cointegration analysis with the approach to modeling based on causal graphs.

Barigozzi et al. (2020) consider I(1) dynamic systems with fewer shocks than variables and that are in this way "singular". Examples of these systems belong to the classes of Dynamic Factor Models (DFM) and DSGE models. They discuss conditions for existence of cointegration and ECM and discuss how the VAR representation can be chosen to have finitely many lags.

Bauer et al. (2020) discuss the system representation of VARMA processes with any integration order at any frequency, using a particular parametrization called the canonical form. They discuss the topological properties of the parametrization, using the cases of $\mathrm{I}(1)$ and $\mathrm{I}(2)$ systems at zero frequency as illustrations. These properties are used to discuss sequences of hypotheses in the $\mathrm{I}(1)$ and $\mathrm{I}(2)$ cases.

Finally, Franchi and Paruolo (2021) discuss the notion of basis of the cointegration space when processes are integrated of any integer order. They show that polynomial cointegration vectors correspond to root functions, for which several results from the literature exists. They show that several polynomial cointegration spaces can be defined for $\mathrm{I}(d)$ systems with $d=2,3, \ldots$, but that a relevant notion (invariant to this choice) is the one of canonical sets of root functions, which act as bases of these spaces. The I(2) case is used to illustrate how some results from the literature can be applied to reduce the number of elements in the canonical set of root functions, i.e., how to make this basis minimal in an appropriate sense.

## 3. Inference

The third set of four papers is concerned with the derivation of new (asymptotic) results for estimation and inference in cointegrated systems (Bernstein and Nielsen 2019; Hansen 2018; Kurita and Nielsen 2019; Li and Bauer 2020).

Hansen (2018) considers GMM estimators for the Reduced Rank Regression model and shows that it is identical to the Maximum Likelihood Estimator under Gaussianity derived in Johansen (1988). This shows that Normality is not needed to motivate the Reduced Rank Regression estimator.

Bernstein and Nielsen (2019) consider the asymptotic distribution of the Likelihood Ratio (LR) test for cointegration rank and of the LR test for known cointegration vectors when the true cointegration rank is lower than the one in the tested hypothesis. They illustrate their results with an analysis of monthly US treasury bonds with one and two year maturity, testing for a stationary yield rate spread.

Kurita and Nielsen (2019) consider partial models with breaks in deterministic terms and Pseudo LR test for the cointegration rank; they derive and tabulate the relevant limit distributions. They illustrate their results with the analysis of partial system of UKGermany log trade balances and the wedge between unit labor costs, conditional on UK and German Gross Domestic Products and the terms of trade.

Li and Bauer (2020) consider estimation in I(2) VAR models when the lag length is chosen as an increasing function of sample size, to allow for VARMA-type data generating processes. Their result are similar to the ones obtained for $\mathrm{I}(2)$ systems with fixed lag-length under appropriate conditions on the growth of the lag-length.

## 4. Modeling and Forecasting

A fourth set of four papers is concerned with modeling and forecasting (Castle et al. 2017; Haldrup and Rosenskjold 2019; Hetland 2018; Hoover 2020).

Hetland (2018) proposes and discusses an extension of the CVAR model called the Stochastic Stationary Root Model. Properties of the process are discussed. Because the likelihood cannot be computed in closed form, a particle filtering approximation is proposed and discussed.

Haldrup and Rosenskjold (2019) consider modeling log death rates by age and time, using US and French mortality tables. They propose a parametric model and fit it with a two step procedure; this allows them to extract four common factors that are later analyzed as a CVAR.

Hoover (2020) discusses the use of CVARs for the analysis of causality links among variables in the form of Directed Acyclical Graphs. An earlier version of this paper generated the problem addressed in Johansen (2019), and the published version of the paper illustrates Johansen (2019)'s results in this context.

Castle et al. (2017) discuss systematic forecast failure, called forediction failure. They propose a step-indicator saturation test to check in advance for invariance of forecast performance to policy changes. A simulation study is used to estimate the potency of this invariance test.

## 5. Applications

A final set of three papers focuses on applications, (Gjelsvik et al. 2020; Goldberg et al. 2020; Lütkepohl and Netšunajev 2018).

Lütkepohl and Netšunajev (2018) study the relationship between the stock market and monetary policy. They consider a CVAR for $\log$ industrial production, $\log$ consumer prices, $\log$ non-energy commodity prices, the $\log$ Euro Stoxx price index and the 3 month Euribor rate. They extend the CVAR model to include a two-states Markov-switching mechanism for the conditional covariance matrix. They use this model to test alternative identification schemes connecting the variables, and produce impulse responses for the chosen specification. For this specification, a contractionary monetary policy shock induces long-lasting (albeit long-run neutral) negative effects on production and on the price level.

Gjelsvik et al. (2020) analyze wage formation in Norway using data from manufacturing, private services and the public sector. They use a partial model of log wages in these three sectors along with the log of the consumer price index, conditionally on a set of other variables. They also allow for broken deterministics and use the critical values derived in Kurita and Nielsen (2019) for cointegration rank determination. They conclude that collective wage negotiations in manufacturing have defined wage norms over the period 1980Q1-2014Q4.

Goldberg et al. (2020) consider the Bilson-Fama regression of future change of the spot exchange rate on the forward premium and find break points for nearly every country. This and further analyses question the widespread view that currency returns are predictable or that developed country markets are less rational.

## 6. Thanks

We would like to thank all contributors to this Special Issue: their willingness to participate and resilience to the editorial review process is what made this Special Issue possible. We also wish to thank Kerry Patterson for asking us to act as Guest Editors for this Special Issue and to Marc Paolella and the Editorial Board of Econometrics for their patience in waiting for it to slowly materialize.

Last but not least, we would like to thank Katarina and Søren for their research, teaching and example. We are indebted to them in many ways, including for their inspiring research in Econometrics.

Conflicts of Interest: The authors declare no conflict of interest. Information and views set out in this paper are those of the authors and do not necessarily reflect the ones of the institutions of affiliation.

## References

Archontakis, Fragiskos, and Rocco Mosconi. 2021. Søren Johansen and Katarina Juselius: A bibliometric analysis of citations through multivariate Bass models. Econometrics 9: 30. [CrossRef]
Barigozzi, Matteo, Marco Lippi, and Matteo Luciani. 2020. Cointegration and error correction mechanisms for singular stochastic vectors. Econometrics 8: 3. [CrossRef]
Bauer, Dietmar, Lukas Matuschek, Patrick de Matos Ribeiro, and Martin Wagner. 2020. A parameterization of models for unit root processes: Structure theory and hypothesis testing. Econometrics 8: 42. [CrossRef]
Bernstein, David H., and Bent Nielsen. 2019. Asymptotic theory for cointegration analysis when the cointegration rank is deficient. Econometrics 7: 6. [CrossRef]
Castle, Jennifer L., David F. Hendry, and Andrew B. Martinez. 2017. Evaluating forecasts, narratives and policy using a test of invariance. Econometrics 5: 39. [CrossRef]
Franchi, Massimo, and Paolo Paruolo. 2021. Cointegration, root functions and minimal bases. Econometrics 9: 31. econometrics9030031. [CrossRef]
Gjelsvik, Marit, Ragnar Nymoen, and Victoria Sparrman. 2020. Cointegration and structure in Norwegian wage-price dynamics. Econometrics 8: 29. [CrossRef]
Goldberg, Michael D., Olesia Kozlova, and Deniz Ozabaci. 2020. Forward rate bias in developed and developing countries: More risky not less rational. Econometrics 8: 43. [CrossRef]
Haldrup, Niels, and Carsten P. T. Rosenskjold. 2019. A parametric factor model of the term structure of mortality. Econometrics 7: 9. [CrossRef]
Hansen, Bruce E. 2018. Johansen's reduced rank estimator is GMM. Econometrics 6: 26. [CrossRef]
Hetland, Andreas. 2018. The stochastic stationary root model. Econometrics 6: 39. [CrossRef]
Hoover, Kevin D. 2020. The discovery of long-run causal order: A preliminary investigation. Econometrics 8: 31. econometrics8030031. [CrossRef]
Johansen, Soren. 1988. Statistical analysis of cointegration vectors. Journal of Economic Dynamics and Control 12: 231-54. [CrossRef]
Johansen, Søren. 2019. Cointegration and adjustment in the $\operatorname{CVAR}(\infty)$ representation of some partially observed CVAR(1) models. Econometrics 7: 2. [CrossRef]
Juselius, Katarina. 2021. Searching for a theory that fits the data: A personal research odyssey. Econometrics 9: 5. econometrics9010005. [CrossRef]
Kurita, Takamitsu, and Bent Nielsen. 2019. Partial cointegrated vector autoregressive models with structural breaks in deterministic terms. Econometrics 7: 42. [CrossRef]
Li, Yuanyuan, and Dietmar Bauer. 2020. Modeling I(2) processes using vector autoregressions where the lag length increases with the sample size. Econometrics 8: 38. [CrossRef]
Lütkepohl, Helmut, and Aleksei Netšunajev. 2018. The relation between monetary policy and the stock market in Europe. Econometrics 6: 36. [CrossRef]

Mosconi, Rocco, and Paolo Paruolo. 2022a. A conversation with Katarina Juselius. Econometrics 10: 20. econometrics10020020. [CrossRef]
Mosconi, Rocco, and Paolo Paruolo. 2022b. A conversation with Søren Johansen. Econometrics 10: 21. [CrossRef]

# A Conversation with Katarina Juselius 

Rocco Mosconi ${ }^{1, *}$ and Paolo Paruolo ${ }^{2}$<br>1 Dipartimento di Ingegneria Gestionale, Politecnico di Milano, Via Lambruschini 4/B, 20156 Milano, Italy<br>2 European Commission, Joint Research Centre, Via E. Fermi 2749, 21027 Ispra, Italy; paolo.paruolo@ec.europa.eu<br>* Correspondence: rocco.mosconi@polimi.it

Citation: Mosconi, Rocco, and Paolo Paruolo. 2022. A Conversation with Katarina Juselius. Econometrics 10: 20. https://doi.org/10.3390/ econometrics10020020

Received: 4 April 2022
Accepted: 6 April 2022
Published: 13 April 2022
Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.


Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/).


#### Abstract

This article was prepared for the Special Issue 'Celebrated Econometricians: Katarina Juselius and Søren Johansen' of Econometrics. It is based on material recorded on 30-31 October 2018 in Copenhagen. It explores Katarina Juselius' research, and discusses inter alia the following issues: equilibrium; short and long-run behaviour; common trends; adjustment; integral and proportional control mechanisms; model building and model comparison; breaks, crisis, learning; univariate versus multivariate modelling; mentoring and the gender gap in Econometrics.


Keywords: cointegration; CVAR; I(1); I(2); common trends; adjustment; breaks; model comparison; gender gap

JEL Classification: C32; B41; C01; C10; C30; C52

## Introduction

On 30-31 October 2018 the authors sat down with Katarina Juselius in Copenhagen to discuss her contributions to Economics and Econometrics. Figure 1 shows photos of Katarina taken on that day; other recent photos are shown in Figure 2. The list of her publications can be found at https://www.Economics.ku.dk/staff/emeriti_kopi/?pure= en/persons/142900. ${ }^{1}$

In the following, frequent reference is made to Vector Autoregressive (VAR) models with Cointegration restrictions, labelled as CVAR, see Juselius (2006)-in particular Part II on the $I(1)$ model and Part $V$ on the $I(2)$ model. In the rest of the article, questions are in bold and answers are in Roman. Text additions are reported between [ ] or in footnotes.

## What do you think of micro-based versus macro-based Macroeconomic models?

I am sceptical to micro-based macro, partly because I have always been critical of the representative agent's approach. In the 1970s there appeared many excellent publications discussing its many unrealistic assumptions on aggregation. But for some reason, the criticism lost steam, the representative agent with rational expectations survived, and micro foundations of macro models became a "must" in academic work.

For me it is puzzling why the criticism did not have a greater impact on the mainstream theory considering that there are so many important aspects on the aggregate economysuch as unemployment, inflation, GDP growth, exchange rate, interest rate, inequality, speculation-that cannot be properly addressed in a representative agent's framework.

Considering all the major problems we face today both in the domestic and the international economy, it is obvious that more than ever we need an empirically well founded macro-based Macroeconomics. While Keynes already laid the foundations for such a theory, the old-fashioned Keynesianism needs of course to be modified to account for what we have learned about expectation formation based on imperfect knowledge/incomplete information, persistent equilibria, coordination failure and much more.


Figure 1. Katarina Juselius, 30 October 2018 in Copenhagen.
Joseph Stiglitz jointly with coauthors has proposed a new approach "disequilibrium Economics", which I think is a promising candidate for such a theory. Whether Stiglitz's disequilibrium Economics will change the direction of Economics is hard to predict, but based on previous experience perhaps one should not be too optimistic. When confronted with serious criticism, the Economics profession has too often responded with silence. For example, after the financial crisis, many methodologically oriented scholars, such as the editors and contributors of the Journal of Economic Methodology, were convinced the time for change had finally come, see Colander et al. (2009).

Numerous books were published addressing the mistakes and the misconceptions leading to the crisis, explaining why things went so wrong, and what could have been done instead. It might seem absurd, but the majority of the profession continued along the same path, modifying some assumptions here and there, but basically continuing as if the financial crisis was just a black swan.

I am often called "heterodox", "not an economist" or just ignored simply because I openly criticize mainstream models for relying on assumptions that do not describe the economic reality well enough. One of our prominent mathematical statisticians, Niels Keiding, once asked me "Why are economists not afraid of empirical data? Medical doctors sure are". While I had no really good answer, I know how hard it has been to raise a serious debate about the great divide between major empirical findings and standard mainstream assumptions, in spite of their important consequences for macroeconomic policy.

How can policy-makers learn about different policy options from a CVAR? Can CVAR answer policy questions?

I believe that policy-makers can primarily benefit from a CVAR analysis because it can improve our understanding of the dynamic transmission mechanisms of basic macroeconomic behaviour. For example, policy-makers facing a problem mostly think of one endogenous variable (the variable of interest) being pushed by a number of exogenous
variables. In practice, the assumed exogenous variables often exhibit strong feed-back effects from changes in the "endogenous" variables.


Figure 2. Katarina Juselius, 3 October 2016 in Milan.
The CVAR does not make use of the endogenous-exogenous dichotomy, but studies the economy as a system allowing for important feed-back effects in all equations of the system. Since policy-makers often are quite conservative in their economic beliefs, a well done CVAR analysis would highlight certain aspects of the model where such beliefs might be incorrect. Hence, a CVAR analysis could help policy makers avoid making bad decisions and, subsequently, to be criticized for them.

Another way a CVAR analysis can be useful is by learning from other countries' experience. For example, Finland, Sweden, and Japan experienced a housing bubble in the early nineties that resembled the more recent house price crisis in 2007. By applying a CVAR analysis to those countries-looking at the crisis mechanisms using the same perspective-policy-makers could have learned more about which policies are likely to work and which are not. They might even have been able to recognize the approaching crisis in time to prevent it.

The usefulness of addressing counter-factual policy questions with the CVAR might be more questionable. Judea Pearl would argue that a model like the CVAR is not appropriate, because the policy variables are set by the policy-maker and are not stochastically generated by the market as the VAR variables are assumed to be. Whether his-mostly theoreticalargument is empirically important is hard to say.

It is certainly the case that a policy variable like the federal funds rate is not behaving like a market determined stochastic variable. But at the same time, a CVAR analysis of the term structure of interest rates-inclusive the fed rate-seems to work reasonably well. But, perhaps one should be a little cautious with the conclusions in such as case.

## What should we learn from crisis periods?

When the economy runs smoothly it doesn't matter much if you have a slightly wrong model, because things work anyway. When you are in a crisis period it matters a lot whether you correctly understand the economic mechanisms and how they work in the economy. The cost of wrong models can then be huge.

The question is, of course, whether it is at all possible to estimate economic mechanisms in a crisis period. For example in the official Danish macro model, the financial crisis is left out altogether with the motivation that it is too extreme to be analyzed econometrically.

I disagree. By experience I know it is possible to get plausible estimates over periods containing a serious crisis.

For example, I have used the CVAR model to address two very serious crisis periods: the house price crisis in Finland (Juselius and Juselius 2014) in the early nineties, and the more recent financial crisis in Greece (Juselius and Dimelis 2019). Both convinced me that it is possible to uncover the destructive forces that unfold during a crisis and that this would help policy-makers to mitigate the worst consequences of a crisis. So in principle I believe it would be a big mistake to leave out a crisis from the sample period.

People have sometimes asked me: "How can you use such periods, which are truly extraordinary, and then expect to find the mechanisms that apply in normal times". This is clearly a relevant question and I may not be able to provide more than a tentative answer: If the sample covers a crisis episode, then one usually needs to apply the $I(2)$ model because it is explicitly specified to account for changes in equilibrium means and/or growth rates. In addition, it is specified to distinguish between levels, changes and acceleration rates, of which the latter is a key aspect of the crisis dynamics.

In normal periods, however, you will observe that acceleration rates are essentially zero. Hence, the acceleration rates take the role of crisis dummies in the I(2) model. However, it should also be acknowledged that the crisis mechanisms of the model may no longer be relevant after the crisis. For example, in the Greek analysis, the crucial crisis mechanism-the strong self-reinforcing mechanism between the bond rate and the unemployment rate-is likely to disappear or at least to change somewhat when the crisis is finally over.

But based on my experience, the main CVAR results seems to hold both for the preand post-crisis period. Perhaps the great variability of the data during a crisis period is also a good thing as it is likely to improve the precision of the estimates.

## The $I(2)$ model is related to the notion of integral and proportional control developed in the 1960s and 1970s. Can the $I(2)$ analysis be useful for understanding the pronounced persistence away from long-run equilibrium relations?

It is hard to come up with any argument to why the $I(2)$ analysis would not be useful, as the $\mathrm{I}(2)$ model is basically designed for integral control. However, the $\mathrm{I}(2)$ model is useful not just in integral control situations, but in a more general setting. If growth rates exhibit persistence and the levels move persistently around long-run trends, then the $I(2)$ model should naturally be the preferred choice.

Also, the $I(2)$ model has also a richer structure than the $I(1)$ model and it has frequently given me insights that I would not have gotten otherwise. This has, in particular, been the case with house and stock prices and the effect of their persistent movements out of equilibrium on the aggregate economy. Without using the I(2) model, I do not think it would be possible to capture the complex mix of error-increasing and error-correcting behaviour in house and stock prices that ultimately led to the financial crisis.

It is quite interesting that economic time-series seem to have become increasingly persistent in the period following financial deregulation. At least I have often found that the I(2) model cannot be rejected based on the trace test for this period. I might have been naive, but I thought this would lead to a greater interest for $\mathrm{I}(2)$ applications. However, when you make a search of " $\mathrm{I}(2)$ " in the Economics literature, you do not find many papers. It is almost as if this fabulously rich model does not exist. But, of course, the I(2) model is more complex than the I(1) model, albeit not more difficult than many other models.

Perhaps, people stay away from the $I(2)$ model because they think that unit roots do not make sense in economic data. And, of course, economic series cannot drift away forever as unit root processes in theory can. If one considers a unit root to be a structural economic parameter, then I agree that neither $\mathrm{I}(2)$ nor $\mathrm{I}(1)$ would make much sense. But, if one thinks, like I do, that unit roots are useful approximations that measure the degree of persistence of economic time series, then it makes a lot of sense.

I believe macroeconomists could do much better by exploiting the richness of the $\mathrm{I}(2)$ model, rather than just ignoring it.

## Do you think the notion of near unit root is crucial for measuring persistence?

I certainly do because near unit root Econometrics provide some powerful tools that help us to uncover important mechanisms that have generated persistence in key economic time-series.

Take for example the unemployment rate, defined as a ratio between zero and one. Because of this, many economists would argue that it is a stationary variable and, hence, should not be modelled as a unit root process. Nevertheless, it is a very persistent near unit root variable for which the largest inverse root [henceforth simply referred to as root] of the autoregressive characteristic polynomial is typically larger than 0.95 .

If you have a sample size of say 80 quarterly observations, you would often not be able to reject the null of a unit root in this case. Many empirical econometricians would, therefore, argue that the unit root approximation is fine as long as the unit root hypothesis cannot be rejected based on the conventional 5\% rule. But the economist would nonetheless (correctly) argue that it is not a structural unit root.

If, instead, we have a sample of 3000 daily observations and an empirical root of 0.99 , then this empirically large root is likely to be rejected as a unit root, even though the degree of persistence is much higher in this case. The $5 \%$ rule has the consequence that the larger the sample size, the easier it is to reject a unit root and vice versa. Hence, sticking to this rule implies that an econometrician would treat a persistent ( 0.9 root) variable as nonstationary and a persistent ( 0.99 root) variable as stationary, whereas an economist would argue that both are stationary independent of the test outcome. Not exactly a situation of clarity!

I hold the pragmatic view that if persistence-for example a long movement away from equilibrium-is an important empirical property of a variable or relation, then we should try to model that property. And one way of doing it is by classifying one's data and relations as $I(0)$, near $I(1)$ and near $I(2)$ and relate them to short run, medium run and long run structures in the data.

For example, a powerful way to uncover the puzzling persistence in unemployment rates is to collect the relevant data and estimate the $I(2)$ model, then find out which other variable(s) are cointegrated with the unemployment rate and how the adjustment takes place in the long, medium and the short run, and which the exogenous forces are. If competently done such a model analysis would help us to understand much more about unemployment persistence and its causes than a conventional model analysis. But near-unit-root Econometrics would probably require a lot more research to offer well worked out procedures for empirical modelling.

I have used this idea to understand the Phillips curve, which has been declared dead numerous times, but still seems to be how policy makers think about unemployment and inflation. The former looks very much like an $I(1)$ series with a small but persistent drift, whereas the latter is almost stationary with a small and persistent drift. That the two series have a different order of persistence explains the lack of empirical support for Phillips curve: inflation rate, being a near $\mathrm{I}(1)$ variable, cannot cointegrate with unemployment rate, being a near I(2) variable. To recover the Phillips curve we need to add at least one more previously omitted (ceteris paribus) variable.

Edmund Phelps argued in his "Structural Slumps" book, see Phelps (1994), that the natural rate of unemployment, rather than a constant, is a function of the real interest rate-possibly also the real exchange rate. I found that the long persistent swings in unemployment rate were cancelled by cointegration with the long-term interest rate implying that they shared a similar persistence and that the residual was cointegrated with the inflation rate. Thus, by exploiting the persistence in the data it was possible to recover the Phillips curve with a Phelpsian natural rate (Juselius and Dimelis 2019; Juselius and Juselius 2014) and, in addition, to learn a lot more about the internal system dynamics and the exogenous forces that had pushed the unemployment rate and the interest rate out of their long-run equilibria.

This way of exploiting the data, I have sometimes called the Sherlock Holmes approach to empirical modelling. By following it you will find results that either support or reject
your priors but you will also find new unexpected results. If you do not sweep the puzzling results under the carpet, but let them rest in your mind, you may very well later come across some new results that put the old puzzles in a new light. These are moments of pure happiness.

At one stage it struck me that I almost always needed to add one or two additional variables to my hypothetical economic relations to achieve stationarity. A systematic feature usually means a common cause. In retrospect, it took me embarrassingly long to realize that the common cause was related to expectations in financial markets formed by imperfect knowledge/incomplete information. Subsequently, I have learnt how crucial the impact of complex feedback dynamics from the financial sector is on the real economy.

## Should inflation and interest rates be treated as stationary, or as I(1) even if they have a long-run equilibrium value?

As I already discussed above, my view of empirical modelling is rather pragmatic, as it has to be because every realistic application is immensely demanding. It is always a struggle to make sense of macroeconomic data relative to the theory supposed to explain it. In this struggle the "perfect" or the "true" easily becomes the enemy of the "good". This applies for sure to the modelling of inflation and interest rates: both of them are crucial for the economy and none of them obey mainstream economic theory.

Like unemployment rate, interest rates can be assumed to be bounded from below by zero (or that is what we previously thought) and from above by some upper limit. Inflation is not necessarily bounded but central banks usually do whatever they can to make it so. Whatever the case, both of them are persistent but differ in degree.

Inflation rates look more like a near $\mathrm{I}(1)$ process, whereas interest rates move in long persistent near $I(2)$ swings around something that could possible be interpreted as a longrun equilibrium value. The question is of course what "long run equilibrium" means if economic relationships do not remain stable over long periods of time. For example, the period before and after financial deregulation describe two completely different regimes. Few equilibrium means remain constant across these two periods.

What is important in my view is that the inverse roots of the characteristic polynomial associated with nominal interest rates often contain a double (near) unit root-or rather one unit root and one near unit root. No theoretical prior would predict such an empirical finding and based on the conventional specific-to-general approach one would probably have swept this puzzling persistence under the carpet. But based on the general-to-specific approach it has been possible to suggest a coherent narrative in which a crucial element is financial market expectations based on imperfect knowledge (Juselius and Stillwagon 2018).

Because the stochastic trend in inflation is persistent of a lower degree than nominal interest rates, one would typically not find cointegration in a bivariate model of inflation and one interest rate and one would have to add at least one more variable. It turns out that by combining inflation with the spread between a short and a long interest rate one usually finds cointegration. This is because the long persistent swing in nominal interest rates are annihilated in the spread which then is cointegrated with the inflation rate. A plausible interpretation is that inflation is cointegrated with expected inflation measured by the spread.

The similarity to the Phillips curve model is quite striking: there we had first to combine unemployment rate with the (long term) interest rate to obtain a stationary cointegration relation for inflation. Thus, the long term interest rate needs to be cointegrated with either the unemployment rate or the short term interest rate to get rid of the persistent swings so that what is left can cointegrate with inflation rate.

Whatever the case, the real interest rate is generally too persistent to be considered stationary even though it is claimed to be so in many empirical papers. Such a claim is often based on a badly specified empirical model and, hence, a sizeable residual error variance that makes statistical testing inefficient. To me, such an analysis represents just a missed opportunity to learn something new.

The CVAR model has parameters related to long-run relations and to short-term adjustment. Other approaches in the literature focus only on the long-term relations, like for example Fully Modified Least Squares. What do you think is the advantage of CVAR over other approaches essentially focusing on the long-run relations?

Believing that a long run relation without the corresponding short-run dynamics is sufficient for understanding an economic problem is like thinking you only need your feet but not your eyes to get to your destination. In a certain sense, bivariate cointegration in a non-stationary world corresponds to correlation in a stationary world. It tells you there is a causal relationship but you need the short-run dynamics to understand the causal links.

There are numerous examples in my publications where it was the short-run dynamics that made me rethink the causal mechanisms of the economic problem. The application to the monetary transmission mechanisms in Denmark-discussed in every detail in my cointegration book (Juselius 2006)-is a good example. I found a stable, completely plausible money demand relation consistent with economic theory, but the estimated short-run dynamics contradicted that theory on essentially all counts. I then spent 10-20 years to understand why.

Another example of why the dynamic adjustment is so crucial is an empirical application to climate change which was done in collaboration with Robert Kaufmann [Robert hereafter], from Boston University (Kaufmann and Juselius 2016). I met Robert at a climate conference in Italy where he introduced me to a fascinating climate data set obtained from ice drilling in Vostok. The data base contained ten climate variables (among others: surface temperature, sea temperature, sea levels, $\mathrm{CO}_{2}$, methane) measured over 400,000 years and based on a frequency of one observation per 1000 years. The dominant feature over this period is the regular occurrence of glacial cycles.

The well known Milankovitch theory was able to associate them with orbital variations such as precession, eccentricity and obliquity of the Earth relative to the Sun. However, the power to explain the glacial cycles with these measures was rather poor and most of the variation in temperature over the glacial cycles remained unexplained. Our purpose was to do better based on a CVAR for 10 climate variables conditional on the Milankovitch orbital variables.

The cointegration results showed that the long-run forces were important, but not as important as the dynamic feedback effects which were able to explain most of the variability. Our conclusion was that if you disregard the short-term adjustment, you will only explain a small part of the glacial cycle phenomena, whereas if you include the short-run feedback you can do much better.

Of course, this type of data and models are not meant to predict what will happen next year, but to learn something about the physics of climate change. For example, our results showed that $\mathrm{CO}_{2}$ was the major determinant of surface temperature in this long period without anthropogenic effects on the climate. The results also showed that the $\mathrm{CO}_{2}$ effect was strengthened by the feedback dynamics. This became strikingly evident when we estimated the effect on temperature from doubling $\mathrm{CO}_{2}$ in the atmosphere.

While most climate models would predict that a doubling leads to an increase of roughly $3^{\circ}$ or $4^{\circ}$ Celsius, our CVAR model predicted an increase of almost $11^{\circ}$ Celsius. That's a difference between something which is a disaster and the end of our civilization. But every time climate scientists update their models it is a little scary to learn that the new version shows that the previous one had again underestimated the effect of $\mathrm{CO}_{2}$ on our climate.

## Did you find I(2) behaviour in that case?

The data display long persistent cycles, between 80,000 and 100,000 years long, which showed up in the model as quite large complex pairs of inverse characteristic roots. The trace test, however, rejected I(2). I believe this was partly because the CVAR is not (yet) designed to handle large cyclical roots close to the unit circle, partly because of the compression of data into 1000 year averages. I would think that if instead we had access to

100 year observations there would be strong evidence of $\mathrm{I}(2)$. This is actually something I still would be keen on studying.

## Your difficulties in publishing the result on $\mathrm{CO}_{2}$ seem to suggest that journals and academia in general are somewhat conservative

"Somewhat conservative" is clearly an understatement. But, science is conservative and for good reasons. One should not jump away from an established path at every whim. What is harder to accept is the stubborn conservatism that is more about protecting one's theoretical stance. I always thought Economics was exceptionally conservative, partly because of its axiomatic foundation which makes it less prone to listen to empirical arguments.

The difficulties with getting our CVAR results published in climate journals suggest that it can also be difficult in physical sciences. I guess that the CVAR methodology may seem difficult and strange the first time you come across it. By now Climate Econometrics has become much more established and it is probably easier to publish papers today using cointegration techniques.

## How can the CVAR methodology affect the learning process in economics?

If adequately done, the CVAR structures the data in economically relevant directions without imposing theory-consistent restrictions on the data prior to testing. By this you give the data the right to speak freely about the underlying mechanisms rather than to force them to speak your favourite story. Macro-data are quite fragile-one realization at each time $t$ from the underlying process-and if you torture them enough they will usually confess. This, I believe, may partly explain the confirmation bias that seems quite prevalent in empirical Economics and which is not how to bring about innovation in learning.

The conventional specific-to-general approach starts with a theory model derived from some basic assumptions which are seldom tested. One example is the assumption of what is endogenous and what is exogenous in the model. Another is the assumption that omitted ceteris paribus variables do not significantly change the obtained results. Both of them tend to be rejected when tested within the CVAR model, and both of them tend to affect the conclusions in a very significant way. If the basic hypotheses are not correct, then the scientific value of the whole modelling analysis is of course questionable, because then it would be impossible to know which results are true empirical findings and which are just reflecting the incorrectly imposed restrictions.

Another example is the assumption of long-run price homogeneity which is an implicit assumption of most economic models. Central banks are mandated only to control CPI inflation, which makes sense under long-run price homogeneity. But over the last $30-40$ years, long-run price homogeneity between CPI prices, house prices and stock prices has consistently been rejected due to the fact that stock prices and house prices have behaved completely differently from CPI prices. Central banks have focused primarily on CPI inflation, and by doing so, contributed to the devastating house and stock price bubbles and a steadily growing inequality in our societies.

I believe these problems could have been avoided if more attention had been paid to the signals in the data which were strong and clear after the financial deregulation in the eighties. But academic professors and policy makers were looking at the data through lenses colored by conventional theory, such as efficient markets, rational expectations and representative agents. Inconsistencies with data evidence were labeled theoretical puzzles and had no consequence for practical policy.

## What can we do to change the status quo?

The question is of course if it is at all possible for empirical Econometrics to break the monopoly of theoretical Economics. While I do not have an answer to this question, I can at least refer to discussions I have had with other scholars.

One possibility is to make use of competitions like in other areas such as architecture. For example, if the government wants to build an opera house, they announce a
competition and whoever has the best project will win the competition. Similarly, if the government want to understand what the mechanisms are behind the soaring house and stock prices in order to avoid a new crisis, they could announce a competition. The team that most convincingly is able to explain past and present crisis mechanisms should win the competition. Of course, I can think of many relevant objections to such competitions, but in any case it might be an important step to bring Macroeconomics closer to empirical reality.

I have discussed these issues many times with David Colander [Dave hereafter], one of the most innovative persons I have ever met. Some years ago he presented a proposal for how to reform university teaching of Economics based on a research oriented line and a more applied line. As the majority of students end up working for governments, research institutes, or institutions like the IMF, the ultimate aim was to offer a better training in how to solve real world problems.

On a practical level, one of Dave's suggestions was a big database into which the government as well as other public and private institutions could upload problems they wanted to be solved. University professors would then be allowed to pick problems related to their area of expertise, work out a proposal for how a research group of professors and students would address the problem, and submit the application to the relevant agency. This would have the advantage of bringing important problems closer to the university and would train students to solve real problems under qualified guidance. I should mention that the above is only a small part of his elaborate proposal which was then available as a written memo.

## Is empirical research in Economics different from Physical Sciences? Do you think that changing the theories starting from evidence in the data is easier there?

Physical sciences tend to agree, to a larger extent than Economics, upon common rules based on which the profession is willing to accept results as being scientifically valid. But when this is said not everyone in physics agrees. For example, when I sometimes discuss the difficulties in social sciences with my son, who is a physicist, he argues that it's more or less the same in his field.

I believe there is a difference in grade in the sense that physical laws are laws in a much stricter sense. Once they have been established, after being suitably tested, they are hard to challenge, whereas economic laws are not "laws" in the same sense, they are much more mental inventions. Hence, one would think that the scientific community would be more willing to modify or change basic assumptions when they appear incompatible with reality.

## In your applied research you address different problems: how do you select your research topics?

The short answer is that my research topics are forced on me by the many "why"s I stumble over in my CVAR analyses. This process started already with my first real economy application to the Danish money demand problem in the late eighties. I was fortunate to find empirical support for a stable, plausible money demand relation.

This was something I was really happy about, but there were other puzzling "why"s associated with the adjustment dynamics. So I decided to study German monetary transmission mechanisms hoping find an answer to my "why"s there. Some of the German results seemed to provide at least partial answers, but then they led to a whole bunch of new "why"s, which I subsequently tried to answer by studying monetary mechanisms in Italy and Spain.

As I was not able to satisfactorily solve the puzzling why's, I turned my attention on the international monetary transmission mechanisms where the purchasing power parity (PPP) and uncovered interest rate parity (UIP) provide the cornerstones. Again, some of the results made sense theoretically, but others raised new "why"s. The most important finding was that PPP needed the UIP to become stationary, indicating that they were inherently tied together.

Michael Goldberg stumbled over my first Journal of Econometrics paper discussing this and told me the results were exactly in accordance with the theory of imperfect knowledge based expectations he and Roman Frydman had worked out. It then dawned on me that many of my "why"s probably had to do with such expectations in financial markets and how they affected the real economy.

Two of the most important variables in the macro economy are the real interest rate and the real exchange rate and both of them exhibited this puzzling persistence. The idea that it was this persistence which had caused the puzzling persistence in unemployment rates suddenly struck me. This was a very important breakthrough in my research. From this stage onwards, I knew the direction.

Another example is a study of foreign aid effectiveness based on 36 African countries, which was commissioned by the UN-WIDER institute. Initially it involved one of my PhD students, but then the project grew and I also became actively involved. As it turned out, among those 36 countries, a few important ones, Tanzania and Ghana, were sticking out in a way that prompted many new "why"s. We picked them out for a much more detailed analysis which subsequently became another research publication. It is the trying to answer the "why"s of one paper that has often led to new papers.

## Let's now discuss model building strategy. can you discuss the role of the deterministic components in the cointegrating vectors? how can structural breaks be distinguished from unit roots?

When I start a new project, I always spend a lot of time examining the graphical display of the relevant data. The first step is to examine the variables in levels and differences searching for features which stick out, such as a change in growth rate or a shift in the level of a variable. At this stage I also check the national economic calendar to identify the time points of major political reforms and interventions, because, in my view, an empirical analysis of a macroeconomic problem is always about combining economic theory with a institutional knowledge.

If I spot a sudden shift in the level of a variable followed by a blip in its difference and it coincides with a known political reform, I will add a shift dummy in the cointegration relations and an impulse dummy in the equations. In the final model I always check whether such a shift dummy is long-run excludable and whether the impulse dummy is statistically significant. The testing is important because a political reform often causes a shift in the equilibrium level of several variables so that the level shift may cancel in the cointegration relations.

While it is good scientific practice to test a prior hypothesis that a break has taken place at a certain known point in time, it is harder to defend a practice where step dummies are added only to be able to accept the stationarity of the variable. For example, as already discussed, unemployment is often found to be a very persistent process with a double near unit root. The trace test frequently concludes that it is not statistically different from an $\mathrm{I}(2)$ process, which can be a problem for a researcher believing it should be stationary.

By introducing sufficiently many deterministic level shifts so that stationarity around the level shifts can be accepted one might be able to solve the dilemma. But, whether you model the variable stochastically with the $I(2)$ model or deterministically with many level shifts, you still need to address the puzzling persistence. I would clearly prefer to model it stochastically unless the breaks coincide with known policy reforms. To introduce breaks for the sole purpose of avoiding the $\mathrm{I}(1)$ or the $\mathrm{I}(2)$ model is not a good practice.

## What about non-normality and dummies?

To assume Gaussian distributions is tempting, because then you have access to a very large tool box. And, because it is extremely demanding to adequately model macroeconomic time-series, you need as many tools as possible. This is because the series are often short, strongly autocorrelated, and subject to regime changes. In addition, macro models have to address path-dependencies, interrelated equations and aggregate behaviour that is typically different in the short, medium and long run. On top of all this inference is based
on a sample where you have just one observation at each time $t$ from an underlying process which seldom is stable over extended periods of time. It is almost a miracle that the VAR model frequently is able to give a satisfactory summary of all this.

However, the assumption that the system is being hit by white noise shocks that cumulate via the dynamics of the process to generate the exogenous trends is a bold one, and an assumption that often needs to be modified.

Empirically, the VAR model is subject to many choices: we choose to study $p$ variables among all the potentially relevant ones and we choose to cut the lag length at a not too large value $k$. In practice, normality is seldom accepted in the first unrestricted version of the VAR model. This is of course no surprise, as the residuals are not really estimates of white noise errors, but instead a summary of everything that has been left out of the model

The effect of omitted variables can to some extent be accounted for by the VAR dynamics. But the effect of policy interventions and reforms are usually part of the residuals. Fortunately, policy events are numerous and their individual effect on the aggregated economy is mostly tiny. Hence, one can use the central limit theorem to justify the normality assumption.

The problem is that the effect of some of the policy events is far from small. For example, financial deregulation had an enormous effect on the economy, value added tax reforms exhibited also a very significant effect. The effect of other extraordinary events such as hurricanes, floods, fires, will often stick out as non-normal residuals. Such extraordinary effects have to be properly controlled for using dummies, or they will bias the VAR estimates. This is because the model will otherwise try to force these big effects onto the $x$ variables.

I usually add dummies one at the time. First the ones I believe have to be there as they are a proxy for a real known event. Then I may add a few more if it is absolutely necessary to achieve residual normality or symmetry. Adding too many dummy variables to the model is generally not a good strategy as large effects are also very informative and dummying them out may destroy the explanatory power of your model.

The graphical display may also show transitory blips in the differenced series, that is a big blip followed by a blip of similar size but of opposite sign. They are typically the consequence of a mistake, sometimes a typing mistake, but mostly a reaction to a market misconception. For example, financial markets often bid up the price of an asset only to realize it was a mistake and the price drops back next period. But because they are symmetrical they affect excess kurtosis and not skewness, which is less serious. I often just leave them as they are. But if the jumps are huge, I usually control for them by a transitory impulse dummy ( $\ldots 0,0,+1,-1,0,0 \ldots)$.

## How do you interpret the results of the trace test? How strictly do you use the 5\% critical value in testing hypotheses?

Some people think that a "rigorous approach" to testing requires a strict adherence to standard rules (such as the $5 \%$ critical value). I have never been an advocate of the $5 \%$ rule, but have always based my choice on the whole range of empirical $p$-values. The $5 \%$ rule is reasonable when you strongly believe in the null hypothesis and, hence, are not willing to give it up unless there is massive evidence against it. Adhering to the $5 \%$ rule is particularly problematic in situations when the econometric null hypothesis does not coincide with the economic null.

The trace test of cointegration rank is a good example. The standard procedure relies on a sequence of tests where you start in the top testing the econometric null hypothesis " $p$ unit roots, that is no cointegration". But this null seldom corresponds to the economic null as it would imply that your preferred economic model has no long run content. If the first null hypothesis is rejected, then you continue until the first time $p-r$ unit roots cannot be rejected. This means that the test procedure is essentially based on the principle of "no prior economic knowledge" regarding the the number of exogenous trends. This is often difficult to justify.

The econometric null is based on the number of unit roots (a simple hypothesis) and a $5 \%$ rule applied to a top-down series of tests will often favour the choice of too many common trends and, hence, too few cointegration relations. This is particularly problematic if your data contains a slowly adjusting economic long-run relation. Given the short samples usually available in Economics, a 5\% trace test will often conclude that a slowly adjusting relation could possibly be a unit root process.

Hence, the top-down test procedure and a (blind) use of 5\% critical values may lead to a rejection of a very plausible economic relation for the sole reason that it has a low mean reversion rate. As if this is not bad enough, treating a stationary relation as a common stochastic trend will also affect your model inference in unknown ways.

To circumvent this problem, I usually start the analysis by asking what is the number of exogenous trends consistent with the economic model in question. I usually test this number using the $5 \%$ rule, but I also check the plausibility of this choice against the closest alternatives, for example based on their trace test statistics and the characteristic roots. When deciding in favour or against adding one more cointegrating relation, I also look at the plausibility of the cointegration relation and the sign and the significance of the corresponding adjustment coefficients.

The most problematic situation is when there is no clear distinction between large and small canonical correlations and, hence, no distinct line between stationary and nonstationary directions. This is often a signal that your information set is not optimally chosen and that some important variables are missing. When in doubt about the right choice of rank I often try to enlarge the information set for example with a potentially important ceteris paribus variable such as the real exchange rate, a variable often ignored in the theory model but extremely important in practice. Surprisingly often this solves the problem.

Another illustration of the misuse of the $5 \%$ rule is the test of long-run exclusion in the CVAR. Here the econometric null is that a variable is not needed in the long run relations. In this case it is hard to argue that the econometric null coincides with the economic null as the variable was chosen precisely because it was considered an important determinant in the long-run relations. To throw it out only because we cannot reject that it might be long-run excludable on the $5 \%$ level seems a little foolish.

The main reason this problem arises is because the econometric null hypothesis is often chosen because of convenience, for example when the econometric null corresponds to a single value whereas the plausible economic null corresponds to a composite hypothesis. Whatever the case, whether you reject or accept a hypothesis, I think you have to openly argue why and then back up your choice with the $p$-value of the test.

## How do you handle, in general, the problem of competing models? Do you like the idea of encompassing proposed by David Hendry?

Yes, I think it is a very useful idea. But I also think it is important to distinguish between encompassing in the econometric sense versus encompassing in the economic sense, even though the two concepts are clearly related. David introduced the concept of encompassing as a way of comparing empirical models. You may consider two models explaining $Y$, one as a function of a subset of $X_{1}$ variables and the other of $X_{2}$ variables. Then you estimate a model for $Y$ as a function of $X_{1}$ and $X_{2}$ and ask which of the two models encompasses the big model.

David Hendry [David hereafter] and Grayham Mizon published the paper "Evaluating Econometric Models by Encompassing the VAR" (Hendry and Mizon 1993) which discussed the general-to-specific principle-which I am very much in favour of-applied to the VAR model as a baseline against which a more specific model should be evaluated. One may say the VAR model provides the econometrician with a set of broad confidence bands within which the empirically relevant model should fall. The advantage of encompassing is that it formalizes a principle for how to weed out models that do not describe the data sufficiently well.

However, the problem of competing models in Economics is even more important as there are many competing schools in Economics but no clear criterion for how to choose between them. Because there is one empirical reality-defined by the relevant data-but several models trying to explain it it seems obvious to discriminate between them by encompassing the CVAR.

I have tried to formalize this idea by the concept of a so called "theory-consistent CVAR scenario", which basically describes a set of testable hypotheses on the pulling and pushing forces in the CVAR model. In short, a scenario specifies a set of empirical regularities that one should find in a CVAR analysis, provided the theoretical assumptions of the economic model were empirically correct. Such a comprehensive testing often reveals a significant discrepancy between theory and empirical evidence.

The crucial question is why the reality differs so much from the theoretical model It is a question that haunted me for many years until I begun to see a systematic pattern in the empirical results. They pointed to some theoretical assumptions associated with expectations in the financial markets that were clearly empirically incorrect but not questioned by the majority of the profession. The scenario analysis made it very explicit where the inconsistencies between theory and empirical evidence were and often helped me to understand why.

But, the formulation of a scenario is no easy task. While I was still actively teaching I used to ask my students to formulate a scenario prior to their econometric analysis, but in most cases it was too difficult without my help. This is a pity, because I am convinced it is a very powerful way to solve the dilemma of competing models in Macroeconomics and to bring macroeconomic models closer to reality.

## Linearity is a common assumption. Do you think it might be important to consider non-linear adjustment?

I consider the CVAR model to be a first order linear approximation to a truly nonlinear world. The question is of course how significant the second order or third order components are. If a first order approximation works reasonably well, then the second or third order components might not be so crucial. But, if the first order approximation works poorly, then it may of course be a good idea to consider for example non-linear adjustment. This could be the case in stock price models where adjustment behaviour is likely to be different in the bull and the bear market. Many people are risk averse and react differently when prices go up than when prices go down, so nonlinearity in the adjustment is likely to be useful in this case.

It is of course much easier to construct a linear model to start with. Take for example the smooth transition model as a very plausible nonlinear adjustment model describing adjustment from one equilibrium level to another. In the linear CVAR model, this can be approximated by a level shift (a step dummy) in the cointegration relations combined with a sufficiently flexible short-run dynamics. In many cases this linear approximation will work almost as well (sometimes better) than the nonlinear alternative.

Another example is the nonlinear model of shifts between stochastically evolving equilibria. These models have been proposed to describe the long-lasting swings we often see in the data. They are typical of variables strongly affected by financial market behaviour such as exchange rates, interest rates, and stock prices which tend to fluctuate between high and low levels. But these stochastically switching equilibrium models can in many cases be more precisely described by the I(2) CVAR model.

As a starting point, I think one could try to approximate potential non-linear effects with the linear $I(1)$ or $I(2)$ model with dummy variables and then exploit the CVAR estimates to develop a better nonlinear model. The difficulty is that the non-linear possibilities are almost infinite which makes it to hard know where to start, unless you have a very clear idea of where in the model the non-linear effects are.

Univariate models, small and large-scale multivariate macro models are all used in applied macroeconomics: how do you think they relate to each other?

Basically, a univariate time-series model of $x_{1}$ is a sub-model of a small-scale multivariate model of $x_{1}, \ldots, x_{k}$, which in turn is a sub-model of a large-scale multivariate model of $x_{1}, \ldots, x_{k}, \ldots, x_{m}$. Hence, one should be able to argue why the smaller model with less information is preferable to a larger model with more information.

It is of course totally acceptable that people can choose between different perspectives when they approach a problem and it may be fully rational to focus on a smaller subset of the relevant information set. What I find to be problematic is the standard use of univariate Dickey-Fuller tests to pre-test the order of integration of each variable of a multivariate model. The absurdity of this becomes obvious when the result of the pre-tests is in conflict with the result of the more informative multivariate tests.

At one stage I became rather frustrated over this lack of coherence. To my great irritation I was often asked by referees to add univariate Dickey-Fuller tests to my papers, which I never did. Also, I consistently demanded any table with such tests to be removed if they had been added by a coauthor. They often reacted with puzzlement: why not calculate the univariate Dickey-Fuller tests? A simple thought experiment explains my concern.

Consider a paper which ultimately is analyzing a CVAR model but starts with a bunch of univariate Dickey-Fuller tests. Imagine now that the univariate pre-tests were placed at the end of the paper. Would this have any effect on the main conclusions of the paper?

I hired a student to find empirical CVAR analyses in papers published in a number of good-ranking journals over a period of 10 years that reported tables with pretesting. In most cases the pretests had no effect whatsoever on the final conclusions. In some cases the pre-tests led the researcher to make incorrect choices such as throwing out a relevant variable that was found to be stationary by the pretests.

To throw out variables as a result of pretesting is of course complete nonsense, because a multivariate model can easily handle a stationary variable but also because a pretested "stationary" variable may not be considered stationary in the multivariate model. This is because what matters is whether a variable corresponds to a unit vector in the cointegration space and this depends on the choice of cointegration rank. If this choice is too small-which is frequently the case-then the pretested "stationary" variable would often be rejected as a unit vector in $\beta$ and the consequence would be a logical inconsistency in the analysis.

The perspective of large-scale macro models is usually different from small-scale models. This is in particular so if by large-scale you mean the large macro models used by finance ministries all over the world. They are typically characterized by a large set of behavioural (and definitional) relationships where the status of variables as endogenous, exogenous and ceteris paribus are assumed a priori and where little attention is given to dynamic feedback effects. As such, it is hard to argue that a small-scale multivariate model is a sub-model of these models as they represent two different approaches to macromodelling. In my book (Juselius 2006) I have proposed a procedure to connect the two.

## Common trends have been shown to be invariant to the extension of the information set. Can this be used to devise a progressive modeling strategy, where results from (one or more) small-scale CVAR models are inputs to a larger scale model?

The unsolved problem here is how to uniquely extract and identify individual common stochastic trends. While it is straightforward to determine the space spanned by the $p-r$ common stochastic trends in a $p$-dimensional CVAR model, it is much more difficult to economically identify these common stochastic trends. Kevin Hoover, Søren and I have worked on this difficult problem for many years with the purpose of solving the problem of long-run causality in economic models.

However, the potential of common trends analysis stretches far beyond this problem. It seems plausible that there are a limited number of common stochastic trends in the world. The invariance of common stochastic trends suggests that we should find linear combinations of these common stochastic trends in small-scale CVAR models. The set
of these extracted common stochastic trends could then be analyzed using cointegration techniques.

Let's say that we have extracted 10 common stochastic trends from a number of small-scale CVAR models and that we find the cointegration rank to be 7. This would be consistent with three fundamental stochastic trends in the economy, for example an inflation trend, a productivity trend, and a financial trend. The problem is, as already said, how to uniquely identify them so that we can put labels on them. I think it is a fantastic research problem.

## What is the role of cross section versus panel data models?

Cross-section models can provide a different perspective on the economy than timeseries models, because they add valuable information about individual characteristics at each point in time unavailable in aggregate time-series data. But the time perspective, such as feedback dynamics, is missing in cross section models.

In panel data models you have the possibility for both perspectives provided you have access to fairly long panel data sets. Personally I think reasonably long consumer panel data sets as we have in Denmark are extremely valuable as they combine the best of the two worlds. But, of course, they can not address all issues of macroeconomic relevance.

An interesting research project that has been on my wish list for a long time is to study the aggregated output from simulated agent-based models to learn more about the connection between micro and macro. For example, would the aggregate behaviour have similar properties in terms of cointegration, adjustment, and feedback dynamics as we usually find in our CVAR models?

## You never used panel cointegration techniques. Is it because you are skeptical about them?

As I already said, I find panel cointegration models based on micro data to be potentially very valuable, but I am much more skeptical about such analyses based on a panel of countries. In my view, countries are individually too different to be merged into the same model structure. In most cases I have come across, the panel analysis is based on so many simplifying assumptions that in the end it is hard to know which of the results are true empirical results and which are due to the simplifying restrictions forced on the data.

For example, one can easily find examples of misuse of country panel data in Development Economics. This is because, for many of these countries, data are only available on an annual basis over a post-colonial period. The quality of data is often low, partly because data collection methods may not be very reliable, partly because observations are missing during periods of war and unrest. This has led many development economists to merge the countries in a panel to get more information out of the data. But as such this is no guarantee that the results to become more reliable; it can easily be the other way around.

To look into this problem, Finn Tarp, Niels Framroze Møller and I started a big project where we studied 36 Sub-Saharan countries regarding the effectiveness of their development aid on GDP growth, investment, private consumption and government expenditure. It was a huge data base and the computer output was almost killing.

Just the initial specification of an adequate VAR for each country was as a major task: first we had to identify the time points for extraordinary events such as wars, military coups, famines, droughts, and floods and then we had to control for them by appropriate dummy variables. Because the individual countries differed a lot, a major task was to classify them into more homogeneous groups.

Niels suggested a first coarse division according to whether aid had a significant long-run effect on GDP or investment, whether aid was exogenous to the system, whether it was purely adjusting to the macro-system, or none of the above. But also within these more homogeneous groups, individual countries differed a lot for example in terms of the magnitude of parameter estimates. We concluded that there were positive and significant effects of aid in almost all countries being studied. This was in stark contrast to panel data
studies published in high ranking journals which showed that foreign aid has had no, or even negative, effect on the growth of GDP and investment.

The lesson seems to be that unless you control properly for extraordinary events and other data problems before pushing the panel data button, you can get basically any result.

So your conclusion is that aid has been effective given some country-specific characteristics. Would it be possible to use these characteristics in a panel data set-up where you include all countries? Could they be a mediator of the effectiveness of aid?

No, I do not really think so. The countries are generally too diverse. As I already mentioned you might be able to use a smaller group as a panel, but not all of them. By studying each country separately, you can identify characteristic features which would be impossible to recognize if they are treated as a homogeneous group.

Take for example a country like Tanzania where Nyerere-the president up to mid-eighties-was a charismatic person with bold visions. The donor countries were generally favourable towards him and Tanzania received a lot of aid, substantially more than any other country in that same period. However, Nyerere believed in a strong currency and used the foreign aid to maintain a fixed exchange rate rather than to improve the development of the country.

Another example is Ghana where a military dictator took over the government in the early seventies. He declared that he had no intention to pay back previous development loans, the perfect recipe for not getting additional loans. As a consequence, the national currency was subject to extreme devaluations followed by hyperinflation.

These are just two examples of the type of heterogeneity you will come across in development countries and they are in no way "black swans". Just to understand the country-specific framework within which the aid is supposed to work requires a lot of work. If you pay attention to all the difficulties that must be solved before you do a panel analysis, the desire to do a panel analysis may evaporate altogether. And if you do all the necessary work, the need for a panel analysis may no longer be so great.

We put a lot of work into this project because of its importance. Rich donor countries give less than $1 \%$ of their GDP in foreign aid to improve the quality of life for very vulnerable people. Even though less than $1 \%$ is not a lot, there are many who would look for a good argument suggesting not to help those who are much worse off. It is unacceptable, in my view, if such an argument is based on too simplified and misleading econometrics.

Our paper was finally published in the Oxford Bulletin (Juselius et al. 2014), after first having been rejected by those Development Economics journals that had published the studies we criticized. One of the Oxford Bulletin referees wrote that he had been of the firm opinion that foreign aid was not contributing to development, but had changed his mind, because-as he wrote-the analysis of our paper was so carefully done that he could not find anything to criticize. I felt very proud.

Can you provide more details on the quality of the data in this research and the consistency across countries of the variables you analyzed?

We used annual data starting from the 60 s, which is when most African countries became independent. But because the 60s was a very volatile transition period we decided to leave out this decade for most countries. The data-consisting of total foreign aid and five key macro variables-were collected from the official data bases, the Penn World Tables and World Development Indicators where data are reported in a reasonably consistent way across countries.

A few countries were excluded due to many missing data points and for two countries we had to add variables to be able to make sense of the results. We were able to keep 36 countries for a detailed CVAR analysis based on roughly 45 annual observations. Every step was carefully reported, but with six variables and only 45 data points it was more or less pointless to apply the recursive tests to check for parameters stability.

Therefore, the parameter estimates should be thought of as representing average effects over the sample period. Despite the shortness of our time series, the data were surprisingly informative, possibly due to their large variation. Still, I think it is plausible that the transmission mechanisms of foreign aid have undergone changes over the last decades, similarly as macroeconomic mechanisms have changed in the industrialized part of the world.

It could, therefore, be quite interesting to extend our study with a more recent data set based on quarterly macro data. For many of the countries such data are available starting from the 90 s. But on the whole I believe the results we obtained from our annual sample were completely plausible, often telling an interesting story about vulnerable economies struggling to find their way out of poverty.

## You said that when the sample is short it is not easy to analyze the stability of the parameters and the possibility of structural breaks: can you elaborate on this?

An interesting example of long-run stability is the Danish money demand relation which is thoroughly analyzed in my book (Juselius 2006). It was my first illustration of Maximum Likelihood cointegration and was published in 1990 in the Oxford Bulletin of Statistics and Economics based on fifteen years of quarterly data from 1972 to 1987 (Johansen and Juselius 1990).

It was a volatile period covering two oil crises, several devaluations of the Danish krona and a far-reaching political decision to deregulate financial movements. A priori there was good reason to suspect that the parameters of the estimated money demand relation would not be totally stable. Even though the recursive stability tests did not signal any problem, the rather short sample of fifteen years made these tests rather uninformative.

At a later stage I updated the data by adding data from 1988 to 1994. To my relief I got essentially the same parameter estimates for the money demand relation as before (Juselius 1998). Based on the extended data, the recursive stability tests were now more informative and they confirmed that the money demand relation was stable.

However, the recursive tests also showed that this was not the case with the first cointegration relation-a partially specified IS relation-which exhibited a complete structural break around mid eighties due to Denmark's financial deregulation. Ironically, the 5\% rule would have selected a non-constant, meaningless, relation and left out the stable and meaningful one, a good illustration of the hazards of blindly using the $5 \%$ rule.

When I started writing my book I decided to use the Danish money demand data as an empirical illustration of the CVAR methodology. A first version of my book was based on the 1994 data set, but in 2004 when the text was more or less finished, I could no longer ignore the fact that the data were rather old. So I updated it once more with additionally 10 years of quarterly observations, now up to 2004.

The first time I run the CVAR with the new data, a whole flock of butterflies fluttered in my stomach. If the empirical results had changed significantly, then I would have had to rewrite large parts of my book. But, fortunately, all major conclusions remained remarkably stable, albeit the estimates changed to some minor extent.

After my retirement in 2014 somebody asked me if the Danish money demand relation was still going strong. So out of curiosity I updated my data once more and found out that the money demand relation was no longer in the data! Adding the period of unprecedented credit expansion that led to the overheated economy ending with the financial crisis, seemed to have destroyed the stability of the money demand relation.

As such it is an interesting finding that prompts the question "why?". Is it because of the exceptionally high house and stock prices in the more extended period compared to the almost zero CPI inflation rate and historically low interest rates have changed the determinants of money demand? Would we be able to recover the old relationship by extending the data with house price and stock price inflation? I would not be too surprised if this was the case.

All this raises an important discussion about the stability of economic mechanisms. The underlying rationale is of course that social norms and behaviour tend to change over time as a consequence of political reforms but also of political views or propaganda, as there nowadays is too much evidence for around the world. However, also economic norms and dogmas are likely to influence behaviour. If economic models show that competition is good and greed is even better then some politician will use it in their propaganda as evidence in favour of their policy.

Of course it would be absolutely fantastic if we had access to powerful econometric tools which could tell us exactly when a structural change has occurred, but I doubt very much this will ever be the case, not even remotely so. Structural change is seldom a black or white event; things change in a much more blurred way. Take for example the overheated economy at the beginning of this century that ended with the financial crisis-the so called long "moderation" period.

If this turns out to be a transitory event, albeit very long-lasting, then the breakdown of the money-demand relation may not represent a structural change. Updating the moneydemand data to the present date might give us back the old parameter estimates. Even though I doubt it very much, it is nonetheless a possibility. In most of my professional life I have struggled with questions like this.

Econometric analysis is fantastic when it helps to make complex structures more transparent, when it forces you to understand puzzling features you would otherwise not have thought about, and when it teaches you to see the world in a new light. But it does not let you escape the fact that it is you who are in charge, it is your judgement and expertise that is a guarantee for the scientific quality of the results.

## You have been mentoring many Ph.d. students and young researchers, like the younger version of the two of us. What did you like or dislike about this? Any forward-looking lessons for other econometricians?

In all these years I have immensely enjoyed guiding students both at the Economics Department in Copenhagen, but also at other departments during our many travels. But to be both a good teacher, a good supervisor and a good researcher at the same time is basically "mission impossible" as long as 24 h a day is a binding restriction. Even though I spent all my time (including late evenings, weekends, and holidays) on these activities, I nevertheless always felt I should have done more.

On top of all this I also had the ambition to engage in the public debate, not to mention obligations to family and friends. So, time was always in short supply and every day was a struggle to meet deadlines and a compromise between everything that needed to be done. It took surprisingly long until my body begun to protest increasingly loudly. In the end it forced me to slow down a little. This is the not-so-good aspect of being an (over)active researcher.

My best teaching memories are without comparison from our many Summer Schools of the Methodology of the Cointegrated VAR. To experience highly motivated, hard-working students willing to give up all other temptations in beautiful Copenhagen only to learn a little more econometrics was a very precious experience and I feel enormously privileged to have had it.

The secret behind this success was that we offered the students a firm theoretical base, a well-worked out guidance for how to apply the theory to realistic problems and a personal guidance of their own individual problems, often a chapter of their PhD thesis. A typical day started with Søren [Johansen] discussing a theoretical aspect of the CVAR (and students came out looking happy and devastated at the same time), then I illustrated the same aspect using the Danish money demand data (students begun to look somewhat more relaxed), and in the early afternoon a teaching assistant explained the same aspect once more based on a new application (students begun to say now they had grasped it).

Finally in the late afternoon, early evening, they had to apply the theory on their own data (students were totally lost, but after competent guidance happiness returned). It was a tough experience, but many students learned immensely in three weeks. One of them said
he had learned more than during three years of full time studies at home. If I should give any lesson for other econometricians, this is the one.

Another extremely good experience was a series of Nordic-later also Europeanworkshops from 1989 to 2000 where we met two, three times a year to discuss ongoing research in the cointegrated VAR model. This was a different way of guiding young researchers-like the two of you-by offering a direct involvement in the research process. It was truly learning-by-doing research. Most of the cointegration results in Søren's book and in my own were developed and intensely discussed in this period. A workshop usually lasted for 3-5 days and we were engaged in discussions every single minute. When the workshop closed I think we were all practically dead.

But I believe we found it enormously exciting. It was a once-in-the-lifetime experience. This is also a lesson I would happily give to other econometricians.

## Is there a "gender gap" in Econometrics? ${ }^{2}$

When I started my academic career as a young econometrician, the gender gap was very large indeed. There were only a few female colleagues at the department and very, very few female professors altogether in Economics. But, even though the gap has become smaller, it has not disappeared.

To some extent, I believe it is a question of a male contra a female culture. The traditional language/jargon in Economics is a male-dominated language foreign to many women. For example, theoretical ideas are formulated in the abstract terms of a "representative agent" who maximizes a well-defined utility function derived from a preference function that often reflects greed.

This way of thinking is not very attractive to many women, who would chose Economics because they are concerned about the huge income gap between industrialized and developing countries, the well-being of their parents, children, friends (not an "agent") and would like to understand why a good friend became unemployed and what to do about it. I think it is quite telling that the two most popular fields among female economists are Labour Economics and Development Economics.

The question is whether the abstract way of formulating Economics is absolutely necessary from a scientific point of view. I find it probematic that trivialities or common sense results are often presented in an almost opaque language which tends to make economic reasoning inaccessible to laymen.

In the book, Chang (2014) "Economics: a user's guide", the well-known Cambridge economist Ha-Joon Chang argues that $95 \%$ of Economics is just common sense, but made more complex by abstract mathematics. His accessible and highly qualified text illustrates this point. On the whole I believe more women would be attracted to research in Economics if one would allow more common sense reasoning and pluralism into the teaching of Economics.

Many times in my teaching, I noticed the cultural difference between male and female students. My male students were often fascinated by the technical aspects, whereas my female students were more excited by the applied aspects. For example, when I demonstrated the derivation of the trace test, the guys were flocking around me after class ended asking questions to the technical aspects. When I illustrated how one could use the technical stuff to ask relevant empirical questions, the female students did the same. They were willing to learn the technical stuff, but mostly because it was necessary for the empirical applications.

The gender gap in publications also reflects a similar difference in attitudes. Many top journals tend to favour "technical" work, partly because it's easier to assess whether a mathematical result is right or wrong than an applied empirical result. But the fact that the editorial boards of top journals are mostly populated by men might also contribute to the gender gap. Notwithstanding today's strong emphasis on empirical work, top journals tend to favour rigorously applied theories and mathematical models which are illustrated with simple examples or alternatively applied to simple problems.

Since many real life problems are much more difficult to formulate using rigorous mathematics they are, therefore, much harder to publish in top journals. When I started teaching the CVAR methodology-which is based on rigorous mathematical statistical principles-I thought it would help female (and male) students to overcome this problem. But it did not work out as I had hoped. The main problem was that the empirical reality seldom supported the rigorously derived economic model.

As I have learnt over and over again, journal editors are not happy to accept a paper reporting results which contradict previously published ones. The consequence was that my PhD students often tried to "sit on two chairs": on one hand they wanted to use the CVAR method in a rigorous way, on the other hand they wished the results to support mainstream economic models. I believe it was still another "mission impossible" and I sometimes regret that I had put them in this situation.

Nowadays there are more female students in Economics than in the past, so things are slowly changing. Many of them are still interested in empirical work often in Labor and Development Economics, but their research is much more related to Microeconometrics than to what I would call disequilibrium Macroeconometrics.

## Did you feel kind of alone in this mainly male environment?

If I feel alone, then it is because I am rather alone in my view about what is important in empirical macroeconomic modelling and how it should be done. Considering all disasters in the world around us, it is obvious to me that we desperately need a much better economic understanding of real world problems, rather than still another model of a toy economy.

I am also aware of the dilemma between rigour and empirical relevance in Economics. I have seen numerous examples of really bad empirical CVAR applications where the data have been read in, the CVAR button has been pushed and meaningless results have been printed out that say nothing useful about our economic reality. While there should be no shortcuts in science and empirical results should be derived in a transparent way obeying accepted scientific rules, I strongly believe there should also be room for informed judgement, what Dave Colander would call "the art of Economics". I also believe this is what a rigorously done CVAR analysis can do for you.

Since I have always been outspoken with my views both in academic forums and in the public debate, I have also got my part of male anger, more nowadays than when I was younger. Perhaps I was more diplomatic then or just more good-looking. Whatever the case, being one of the very few female economists was not just negative, I probably did not have to fight as hard for attention as a comparable male econometrician.

But the fact that my research has received a lot of interest among econometricians, economic methodologists and the public is something I value very highly. Many of my absolutely best and most valued colleagues and friends are male economists or econometricians, as for example Søren, the guest editors and the contributors of this wonderful Special Issue. So, on the whole I have been very fortunate in my professional life.

Author Contributions: Both authors have contributed to all phases of the work. All authors have read and agreed to the published version of the manuscript.
Funding: No external funding was received by the second author.
Institutional Review Board Statement: Not applicable.
Informed Consent Statement: Not applicable.
Data Availability Statement: Not applicable.
Acknowledgments: Authors gratefully acknowledge useful comments and correction from the interviewee on the draft of the article.

Conflicts of Interest: The authors declare no conflict of interest. Information and views set out in this paper are those of the authors and do not necessarily reflect the ones of the institutions of affiliation.

## Notes

Accessed on 8 April 2022.
2 See Card et al. (2021).

## References

Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri. 2021. Gender Differences in Peer Recognition by Economists. Technical Report 28942. Cambridge: National Bureau of Economic Research. [CrossRef]
Chang, Ha-Joon. 2014. Economics: A User's Guide. Torrance: Pelican.
Colander, David, Michael Goldberg, Armin Haas, Katarina Juselius, Alan Kirman, Thomas Lux, and Birgitte Sloth. 2009. The financial crisis and the systemic failure of academic economics. Critical Review 21: 249-267. [CrossRef]
Hendry, David F., and Graham E. Mizon. 1993. Evaluating dynamic econometric models by encompassing the VAR. In Models, Methods and Applications of Econometrics. Edited by Peter C.B. Phillips. Oxford: Blackwell, pp. 272-300.
Johansen, Søren, and Katarina Juselius. 1990. Maximum likelihood estimation and inference on cointegration-With applications to the demand for money. Oxford Bulletin of Economics and Statistics 52: 169-210. [CrossRef]
Juselius, Katarina. 1998. A structured VAR for Denmark under changing monetary regimes. Journal of Business E Economic Statistics 16: 400-11. [CrossRef]
Juselius, Katarina. 2006. The Cointegrated VAR Model: Methodology and Applications. Oxford: Oxford University Press.
Juselius, Katarina, and Sophia Dimelis. 2019. The Greek crisis: A story of self-reinforcing feedback mechanisms. Economics: The Open-Access, Open-Assessment E-Journal 13: 1-22. [CrossRef]
Juselius, Katarina, and Mikael Juselius. 2014. Balance sheet recessions and time-varying coefficients in a Phillips curve relationship: An application to Finnish data. In Essays in Nonlinear Time Series Econometrics. Edited by Niels Haldrup, Mika Meitz and Pentti Saikkonen. Oxford: Oxford University Press.
Juselius, Katarina, Niels Framroze Møller, and Finn Tarp. 2014. The long-run impact of foreign aid in 36 African countries: Insights from multivariate time series analysis. Oxford Bulletin of Economics and Statistics 76: 153-84. [CrossRef]
Juselius, Katarina, and Josh R. Stillwagon. 2018. Are outcomes driving expectations or the other way around? An I(2) CVAR analysis of interest rate expectations in the dollar/pound market. Journal of International Money and Finance 83: 93-105. [CrossRef]
Kaufmann, Robert K., and Katarina Juselius. 2016. Testing competing forms of the Milankovitch hypothesis: A multivariate approach. Paleoceanography 31: 286-97. [CrossRef]
Phelps, Edmund. 1994. Structural Slumps. Princeton: Princeton University Press.

Editorial

# A Conversation with Søren Johansen 

Rocco Mosconi ${ }^{1}$ and Paolo Paruolo ${ }^{2, *}$

Citation: Mosconi, Rocco, and Paolo Paruolo. 2022. A Conversation with Søren Johansen. Econometrics 10: 21.
https: / /doi.org/10.3390/
econometrics10020021

Received: 4 April 2022
Accepted: 6 April 2022
Published: 13 April 2022
Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.


Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/).

1 Dipartimento di Ingegneria Gestionale, Politecnico di Milano, Via Lambruschini 4/B, 20156 Milano, Italy; rocco.mosconi@polimi.it
2 European Commission, Joint Research Centre (JRC), Via E. Fermi 2749, 21027 Ispra, Italy

* Correspondence: paolo.paruolo@ec.europa.eu


#### Abstract

This article was prepared for the Special Issue "Celebrated Econometricians: Katarina Juselius and Søren Johansen" of Econometrics. It is based on material recorded on 30 October 2018 in Copenhagen. It explores Søren Johansen's research, and discusses inter alia the following issues: estimation and inference for nonstationary time series of the $I(1), I(2)$ and fractional cointegration types; survival analysis; statistical modelling; likelihood; econometric methodology; the teaching and practice of Statistics and Econometrics.


Keywords: cointegration; fractional (co-)integration; statistical model; survival analysis; VAR; I(1); I(2)
JEL Classification: C32; B41; C01; C10; C30; C52

## Introduction

On 30 October 2018 the authors sat down with Søren Johansen in Copenhagen to discuss his wide-ranging contributions to science, with a focus on Econometrics. Figure 1 reports a photo of Søren taken on the day of the conversation; other recent photos are reported in Figure 2. The list of his publications can be found at the following link: http:/ / web.math.ku.dk/~sjo/. ${ }^{1}$

In the following, frequent reference is made to vector autoregressive (VAR) equations of order $k$ for a $p \times 1$ vector process, $X_{t}$, for $t=1, \ldots, T$, of the following form:

$$
\begin{equation*}
\Delta X_{t}=\Pi X_{t-1}+\sum_{i=1}^{k-1} \Gamma_{i} \Delta X_{t-i}+\varepsilon_{t} \tag{1}
\end{equation*}
$$

where $\Pi$ and $\Gamma_{i}$ are $p \times p$ matrices, and $\Delta=1-L$ and $L$ are the difference and the lag operators, respectively.

Various models of interest in cointegration are special cases of (1), in particular the cointegrated VAR (CVAR), defined by restricting $\Pi$ in (1) to have reduced rank, i.e., $\Pi=\alpha \beta^{\prime}$ with $\alpha$ and $\beta$ of dimension $p \times r, r<p$. Another matrix of interest is the $p \times p$ matrix $\Gamma=I-\sum_{i=1}^{k-1} \Gamma_{i}$, see Johansen (1996, chp. 4) for further reference. For any matrix $\alpha, \alpha_{\perp}$ indicates a basis of the orthogonal complement to the span of $\alpha$; this orthogonal complement is the set of all vectors orthogonal to any linear combinations of the column vectors in $\alpha$.

In the rest of the article, questions are in bold and answers are in Roman. Text additions are reported between [ ] or in footnotes. Whenever a working paper was later published, only the published paper is referenced. The sequence of topics covered in the conversation is as follows: cointegration and identification; survival analysis and convexity; model specification.


Figure 1. Søren Johansen, 30 October 2018 in Copenhagen.

## What is your current research about?

I worked on several projects. With Bent Nielsen [referred to as Bent hereafter] I have studied some algorithms and estimators in robust statistics including M-estimators, see Johansen and Nielsen (2019), and with Morten Ørregaard Nielsen [referred to as Morten hereafter] I have worked on fractional cointegration and other topics in cointegration, see for instance the paper on a general formulation for deterministic terms in a cointegrated VAR model Johansen and Nielsen (2018).

I have collaborated with Kevin Hoover on the analysis of some causal graphs, and just written a paper for this Special Issue (Johansen 2019) on the problem that for a CVAR the marginal distribution of some of the variables is in general an infinite order CVAR, and one would like to know what the $\alpha$ coefficients in the marginal model are.

I have also recently worked with Eric Hillebrand and Torben Schmith (Hillebrand et al. 2020) on a cointegration analysis of the time series of temperature and sea level, for the Special Issue for David Hendry in the same journal. We compare the estimates for a number of different models, when the sample is extended. There has been a growing interest in using cointegration analysis in the analysis of climate data, but the models have to be built carefully taking into account the physical models in this area of science.

The notion of cointegrating space was implicit in Engle and Granger's 1987 paper. You mentioned it explicitly in a paper of yours in $1988 .^{2}$ Could you elaborate on this?

When you realize that linear combinations of cointegrating vectors are again cointegrating, it is natural to formulate this by saying that the cointegrating vectors form a vector space. That of course implies that you have to call the zero vector "cointegrating", even if there are no variables involved. Moreover a unit vector is also cointegrating, even though only one variable is involved. I sometimes try to avoid the word "cointegration", which obviously has connotations to more then just one variable, and just talk about stationary linear combinations.

This lack of acceptance, that a cointegrating vector can be a unit vector, is probably what leads to the basic misunderstanding that almost every applied paper with cointegration starts with testing for unit roots with univariate Dickey-Fuller tests, probably with the consequence that stationary variables will not be included in the rest of the analysis. It is, I think, quite clear that analysing the stationarity of individual vectors in a multivariate framework by testing for a unit vector in the cointegrating space is more efficient than trying to exclude variables from the outset for irrelevant reasons.


Figure 2. Søren Johansen: (a) 24 February 2016 in Copenhagen; (b) 3 October 2016 in Milan.
Going back to the cointegrating space, it is a natural concept in the following sense. The individual cointegrating relations are not identified, and one has to use restrictions from economic theory to identify them. But the cointegrating space itself is identified, thus it is the natural object to estimate from the data in the first analysis.

Hence the cointegrating space is a formulation of what you can estimate without having any special knowledge (i.e., identifying restrictions) about the individual cointegrating relations. The span of $\beta$ (which is the cointegrating space) is therefore a useful notion.

Estimation and testing for cointegration are sometimes addressed in the framework of a single equation.

When estimating a cointegrating relation using regression, you get consistent estimates, but not valid $t$-statistics. Robert Engle [referred to as Rob hereafter] worked out a threestep Engle-Granger regression which was efficient, see Engle and Yoo (1991). Later Peter Phillips (1995) introduced the fully modified regression estimator, where the long-run variance is first estimated and then used to correct the variables, followed by a regression of the modified variables. If there are more cointegrating relations in the system, and you only estimate one, you will pick up the one with the smallest residual variance. It is, however, a single equation analysis and not a system analysis, as I think one should try to do.

## How were your discussions on cointegration with the group in San Diego?

My contact with the econometric group in San Diego started when I met Katarina Juselius [referred to as Katarina hereafter]. She had met David Hendry [referred to as David hereafter] while on sabbatical at London School of Economics in 1979. She was one of the first to use PcGive. ${ }^{3}$ Rob was visiting David in those days. That meant that in 1985, when we went for a month to San Diego, we met Clive Granger [referred to as Clive hereafter], Helmut Lütkepohl and Timo Teräsvirta. So when I started to work on cointegration we knew all the right people.

We were well received and discussed all the time. Clive was not so interested in the technicalities I was working on, but was happy to see that his ideas were used. Rob, however, was more interested in the details. When we met a few years later at the 1987 European Meeting of the Econometric Society in Copenhagen, he spent most of his lecture talking about my results, which is the best welcome one can receive.

So I was certainly in the inner group from the beginning. In 1989, we spent three months in San Diego with Clive, Rob, David, Timo Teräsvirta and Tony Hall. That was really a fantastic time we had. There was not any real collaboration, but lots of lectures and discussions.

I later collaborated with David on the algorithms for indicator saturation he had suggested. His idea was to have as many dummy regressors a you have observations. By including first one half and then the other half you get a regression estimator, and we found the asymptotic properties of that, see Santos et al. (2008).

Later I continued to work on this with Bent, see Johansen and Nielsen (2009); that lead to a number of papers on algorithms, rather than likelihood methods. We analysed outlier detection algorithms and published it in Johansen and Nielsen (2016b), and a paper on the forward search, Johansen and Nielsen (2016a).

## How was cointegration being discussed in the early days?

Clive in Engle and Granger (1987) was the first to suggest that economic processes could be linear combinations of stationary as well as nonstationary processes, and thereby allowing for the possibility that linear combinations could eliminate the nonstationary components. That point of view was a bit difficult to accept for those who worked with economic data. I think the general attitude was that each macroeconomic series had its own nonstationary component.

In Engle and Granger (1987) they modelled the multivariate process as a moving average process with a non-invertible impact matrix, and they showed the surprising result that this "non-invertible" system could in fact be inverted to an autoregressive model (with infinite lag length). Thus a very simple relation was made to the error correction (or equilibrium correction) models studied and used at London School of Economics.

David was analysing macroeconomic data like income and consumption using the equilibrium correcting models, see Davidson et al. (1978). He realized very early that some of the results derived from the model looked more reasonable if you include the spread between income and consumption (for instance) rather than the levels of both. He did not connect it to the presence of nonstationarity.

One of the first applications of the ideas of cointegration was Campbell and Shiller (1987), who studied the present value model in the context of a cointegrating relation in a VAR. The first application of the CVAR methodology was Johansen and Juselius (1990). Here the model is explained in great detail, and it is shown how to test hypotheses on the parameters. Everything is exemplified by data from the Danish and Finnish economies.

Another early paper of the CVAR was an analysis of interest rates, assumed to be nonstationary, while still the spreads could be stationary, as discussed in Hall et al. (1992). These papers contain examples where one can see directly the use and interpretation of cointegration.

## How did you start thinking about identification of cointegrating vectors?

The identification problem for cointegrating relations is the same as the identification problem discussed by the Cowles Commission, who modelled simultaneous equations for macro variables and needed to impose linear restrictions to identify the equations. We were doing something similar, but trying to model nonstationary variables allowing for linear cointegrating relations, and we needed linear restrictions on the cointegrating coefficients $\beta$ in (1) in order to distinguish and interpret them.

Then one can use the Wald condition for identification, which requires that the matrix you get by applying the restrictions of one equation to the parameters of the other linear equations should have full rank $r-1$, see e.g., Fisher (1966) Theorem 2.3.1. This condition, however, contains the Data Generating Process (DGP) parameter values. This implies that the rank condition cannot be checked in practice, because the DGP is unknown. I asked David what he would do, and he said that he checks the Wald rank condition using
uniform random numbers on the interval $[0,1]$ instead of the true unknown parameters. This approach inspired me to look for the mathematics behind this.

## How did you derive the explicit rank conditions for identification?

For simultaneous equations, the restrictions $R_{i}$ imposed on the parameters $\theta_{i}$ of equation $i, R_{i}^{\prime} \theta_{i}=0$, define also a parametrization using the orthogonal complement $H_{i}=R_{i \perp}$ and the parameter is $\theta_{i}=H_{i} \phi_{i}$. The classical Wald result is that if $\theta$ denotes the matrix of coefficients of the DGP for the whole system, then $\theta$ is identified if and only if the rank of the matrix $R_{i}^{\prime} \theta$ is $r-1$ for all $i$.

I realized soon that I should apply the restrictions not to the parameters but to the parametrizations as given by the orthogonal complements of the restrictions, and the Wald condition can be formulated as the condition rank $\left(R_{i}^{\prime}\left(H_{i_{1}}, \ldots, H_{i_{k}}\right)\right) \geq k$ for any set of $k$ indices not containing $i$. This condition does not involve the DGP values and, if identification breaks down, it can be used to find which restrictions are ruining identification.

I reformulated the problem many times and my attention was drawn to operations research, so I asked Laurence Wolsey, when I was visiting the University of Louvain, who suggested the connection to Hall's Theorem (for zero restrictions) and Rado's Theorem (for general linear restrictions), see Welsh (1976). The results are published in Johansen (1995a).

The solution found was incorporated in the computer programs we used when we developed the theory for cointegration analysis. With a moderate amount of equations, the results can be useful to modify the restrictions if they are not identifying, by finding out which restrictions cause the failure of identification.

The value added of this result is the insight: we understand the problem better now, and finding where these conditions fail can help you reformulate better exclusion restrictions. Katarina has developed an intuition for using these conditions, which I do not have. You need to have economic insight to see what is interesting here; for me, it is a nice mathematical result.

I also discussed the result with Rob and he said that it's interesting to see the identification problem being brought back into Econometrics. After Sims' work, identification of systems of equations had been sort of abandoned, because in Sims' words, you had "incredible sets of restrictions".

## You introduced reduced rank regression in cointegration. How did this come about?

In mathematics, you reformulate a problem until you find a solution, and then you sometimes find that someone else has solved the problem-this is what happened with reduced rank regression in cointegration, which I worked out as the Gaussian maximum likelihood estimation in the cointegrated VAR model.

When I first presented the results-later published in Johansen (1988b)-at the European Meeting of the Econometric Society in 1987 in Copenhagen, I was fortunate to have Helmut Lütkepohl in the audience who said: "isn't that just reduced rank regression?". This helped me include references to Anderson (1951), Velu et al. (1986) and to the working paper version of Ahn and Reinsel (1990). Finally, reduced rank regression is also used in limited information calculations, which can be found in many textbooks.

I used Gaussian maximum likelihood to derive the reduced rank estimator, but Bruce Hansen in this Special Issue, Hansen (2018), makes an interesting point, namely that reduced rank regression is a GMM-type estimator, not only a Gaussian Maximum Likelihood solution.

Finally, my analysis revealed a kind of duality between $\beta$ and $\alpha_{\perp}$ which can be exploited to see how many models can be analysed by reduced rank regression. As summarized in my book (Johansen 1996) reduced rank regression can be used to estimate quite a number of different submodels, with linear restrictions on $\beta$ and /or $\alpha$ and allowing different types of deterministic terms. But of course it is easy to find sub-models, where one has to use iterative methods to find the maximum likelihood estimator.

## How did you start working on Granger-type representation theorems?

In 1985 I was shown by Katarina the original working paper by Clive before it was published; this was when I started working on cointegration. I started with an autoregressive representation of a process, and found its moving average representation that Clive used as the starting point. I find that a more satisfactory formulation, trying to understand the structure of what he was working on, and I produced the paper on the mathematical structure, Johansen (1988a).

I was looking for something simple in the very complicated general case with processes integrated of any integer order, and I settled to focus on what I called the "balanced case", that is a relation between variables that are all differenced the same number of times. The balanced case is very simple, and was a way of avoiding a too complicated structure. However, I was focusing on the wrong case, because it is the unbalanced case which is of importance in the I(2) model.

The mathematical structure paper, however, contains "the non-I(2) condition" (see Theorem 2.5 there), which states that $\alpha_{\perp}^{\prime} \Gamma \beta_{\perp}$ need to be full rank in I(1) VAR systems in (1) with $\Pi=\alpha \beta^{\prime}$. That came out as just one small result in this large paper, but that was the important result which was missed in the Engle and Granger (1987) paper.

## This links to the I(2) model and its development.

In 1990 Katarina obtained a grant from the Joint Committee of the Nordic Social Sciences Research Council. The purpose was to bring together Ph.D. students in Econometrics together with people working in private and public institutions in the Nordic Countries to teach and develop the theory and the applications of cointegration. We had two to three workshops a year for 6 or 7 years. The work we did is documented in Juselius (1994) [see Figure 3].

In the beginning, Katarina and I would be doing the teaching and the rest would listen, but eventually they took over and presented various applications. It was extremely inspiring to have discussions on which direction the theory should be developed. One such direction was the I(2) model, and I remember coming to a meeting in Norway with the first computer programs for the analysis of the I(2) model on Katarina's portable Toshiba computer with a liquid crystal screen.

It was a very inspiring system we had, where questions would be raised at one meeting and I would then provide the answers at the next meeting half a year later. Identification was discussed, $I(2)$ was discussed, and computer programs were developed, and people would try them out. I kept the role as the "mathematician" in the group all the time and decided early on that I would not try to go into the Economics.

## Which $I(2)$ results came first?

The I(2) model was developed because we needed the results for the empirical analyses in the group, and the first result was the representation theorem, Johansen (1992). This contained the condition for the process generated by the CVAR to have solutions which are $\mathrm{I}(2)$, generalizing "the non- $\mathrm{I}(2)$ condition" to "the non- $\mathrm{I}(3)$ condition".

The next problem I took up was a systematic way of testing for the ranks of the cointegrating spaces, which I formulated as a two stage analysis for ranks, Johansen (1995b). This problem was taken up by Anders Rahbek and Heino Bohn Nielsen who took over and analysed the likelihood ratio test for the cointegration ranks, Nielsen and Rahbek (2007).

The likelihood analysis for the maximum likelihood estimation of the parameters is from Johansen (1997). When I developed the I(2) model, I realized that the balanced case is not the interesting one. You need relationships for the $I(2)$ processes of the type $\beta^{\prime} X_{t}+\varphi^{\prime} \Delta X_{t}$ to reach stationarity, and this is the so-called "multi-cointegration" notion.

I realized from the very beginning that Clive's structure with the reduced rank matrix in the autoregressive model $\Pi=\alpha \beta^{\prime}$ in (1) is an interesting structure. So one wants to see how one can generalize it. This of course can be done in many ways but the collaboration with Katarina on the examples was very inspiring.


Figure 3. Front and back cover of Juselius (1994), vol. I (of IV). Areas in black indicate sites where the Nordic workshops took place between 1990 and 1993.

One such example is to take two $\log$ price indices $p_{i t}, i=1,2$, where each one is $\mathrm{I}(2)$, but $p_{1 t}-p_{2 t}$ is $\mathrm{I}(1)$; one could then have that $p_{1 t}-p_{2 t}+\varphi \Delta p_{1 t}$ comes down to stationarity, where $\Delta p_{1 t}$ is an inflation rate and $\varphi$ is some coefficient. She pointed out that the important part of the $I(2)$ model was that it allowed for the combination of levels and differences in a single equation, and this is exactly the unbalanced case. In order to understand this I needed to go back and first work out the representation theory, and then start on the statistical analysis.

## What asymptotic results did you derive first?

The asymptotics for the rank test in the $\mathrm{I}(1)$ model came first. I attended a meeting at Cornell in 1987, where I presented the paper on the mathematical structure of error correction models (Johansen 1988a). I included one result on inference, the test for rank. For that you need to understand the likelihood function and the limits of the score and information. I could find many of the results, but the limit distribution of the test for rank kept being very complicated.

At the conference I met Yoon Park who pointed out that the limit distributions had many nuisance parameters, and that one could try to get rid of them. This prompted me to work through the night to see if the nuisance parameters would disappear in the limit. I succeeded and could present the results in my lecture the next day.

So the mathematical structure paper Johansen (1988a) had the rank test in it and its limit distribution, see Section 5 there. The most useful result was that the limit distribution of the test for rank $r$ is the same as if you test that $\Pi=0$ in the CVAR with one lag and $p-r$ dimensions, that is, a multivariate setup for the analogue of the Dickey-Fuller test.

The limit distribution for the rank test with Brownian motions is something I always showed as a nice result when I lectured on it, but it is in a sense not so useful for analysis, because we don't know its mean, variance, or quantiles. So to produce the tables of the asymptotic distribution you must go back to the eigenvalue problem with random walks and then simulate the distribution for a sufficiently large value of $T$.

I think, the next result I worked on was the limit distribution for $\hat{\beta}$. It was derived using the techniques that Peter Phillips had developed, see Phillips (1986). He had picked
the right results on Brownian motion from probability and used them to analyse various estimators, and I could simply use the same techniques.

Ted Anderson's reduced rank regression, Peter Phillips' Brownian motions, Phil Howlett's results (about which I found out much later) on the non-I(2) condition (Howlett 1982) were all fundamental to my work, but the reason that I could exploit all these methods and results was my basic training in probability theory, and I am very grateful for the course Patrick Billingsley gave in Copenhagen in 1964-1965.

## What are recent related results that you find interesting?

The paper by Onatski and Wang (Onatski and Wang 2018) has some very nice results. They consider a multivariate Dickey Fuller test, testing that $\Pi=0$ in the VAR in (1). They let the dimension $p$ of the system go to infinity proportionally to the number of observations $T$, and they get an explicit limit distribution. This is based on results on the eigenvalues of matrices of i.i.d. observations in large dimensions, which has been studied in Mathematics and Statistics. Onatski and Wang have an explicit expression for the limit distribution of the multivariate Dickey Fuller test, called the Wachter distribution.

They refer to the paper Johansen et al. (2005) where we do the simulations to discuss Bartlett's correction. Part of that is simply simulating the multivariate Dickey Fuller test for different dimensions and different $p$. And they show that their asymptotic formula fits nicely with our simulations. Extensions to cases with deterministic terms and breaks, and the ones for rank different from 0 should be carefully considered.

## Tell us about your contribution to fractional cointegration.

Morten wrote his thesis on fractional processes 2003 at Aarhus University, and I was asked to sit on his committee. Some years later I had formulated and proved the Granger representation theorem for the fractional CVAR (FCVAR) in Johansen (2008), where the solution is a multivariate fractional process of order $d$, which cointegrated to order $d-b$. We decided to extend the statistical analysis from the usual CVAR to this new model for fractional processes.

The fractional processes had of course been studied by many authors including Peter Robinson and his coauthors, like Marinucci, Hualde and many others. There are therefore many results on the stochastic behaviour of fractional process on which we could build our statistical analysis.

The topic had mostly been dealt with by analyzing various regression estimators and spectral density estimators, where high level assumptions are made on the data generating process. I thought it would be interesting to build a statistical model, where the solution is the fractional process, so one can check the assumptions for the model.

We had the natural framework in the VAR model, and we just needed to modify the definition of differences and work out properties of the solution. From such a model one could then produce (likelihood) estimators and tests, and mimic the development of the CVAR.

We decided, however, to start with the univariate case, simply to get used to the analysis and evaluation of fractional coefficients. We published that in Johansen and Nielsen (2010), and our main results on the FCVAR, that is the fractional CVAR, are in Johansen and Nielsen (2012).

It helped the analysis that for given fractional parameters $b$ and $d$, the FCVAR model can be estimated by reduced rank regression. We found that inference on the cointegrating relations is mixed Gaussian, but now of course using the fractional Brownian motion, so basically all the usual results carry over from the CVAR.

We are currently working on a model where each variable is allowed its own fractional order, yet after suitable differencing, we can formulate the phenomenon of cointegration. The analysis is quite hard with some surprising results. It turns out that inference is asymptotically mixed Gaussian both for the cointegrating coefficients, but also for the difference in fractional order.

## For fractional cointegration, you appear to be attracted more by the beauty of the model and the complexity of the problem, rather than the applications. Is this the case?

You are absolutely right. There is not a long tradition for the application of fractional processes in Econometrics, even though some of the examples are financial data, where for instance log volatility shows clear sign of fractionality and so do interest rates when measured at high frequency, see Andersen et al. (2001).

Clive and also other people have tried to show that fractionality can be generated by aggregation. Granger (1980) takes a set of $\operatorname{AR}(1)$ autoregressive coefficients with a cross-sectional beta distribution between -1 and +1 ; then integrating (aggregating) he gets fractionality of the aggregate. However, if you choose some other distribution, you do not get fractionality.

As another source of fractionality, Parke (1999) considered a sum of white noise components $\varepsilon_{t}$ which are dropped from the sum with some given probability. If you choose some specific waiting time distribution, you obtain the spectrum or auto-covariance function of a fractional process. There is also another result by Diebold and Inoue (2001) who show that a Markov switching model generates fractionality. Still we lack economic questions that lead to fractionality.

I read about an interesting biological study of the signal from the brain to the fingers. The experiment was set up with a person tapping the rhythm of a metronome with a finger. After some time the metronome was stopped and the person had to continue tapping the same rhythm for a quarter of an hour. The idea was that the brain has a memory of the rhythm, but it has to send a signal to the fingers, and that is transmitted with an error. The biologist used a long memory process (plus a short memory noise) to model the signal.

## Have you ever discussed fractional cointegration with Katarina?

No, she refuses to have anything to do with it, because she is interested in Macroeconomics. She feels strongly that the little extra you could learn by understanding long memory, would not be very interesting in Macroeconomics. It will also take her interest away from the essence, and I think she's right. In finance, something else happens. Here you have high frequency data, and that seems a better place for the fractional ideas.

## Tell us about your contributions in survival analysis.

I spent many years on developing the mathematical theory of product integration, which I used in my work on Markov chains, Johansen and Ramsey (1979). I later collaborated with Richard Gill on a systematic theory of product integration and its application to Statistics, Gill and Johansen (1990). The interest in the statistical application of product integration came when I met Odd Aalen in Copenhagen. He had just finished a Ph.D. on the theory of survival analysis using counting processes, with Lucien Le Cam from Berkeley, and was spending some time in Copenhagen.

Towards the end of his stay, he presented me with a good problem: he asked me if I could find the asymptotic distribution of the Kaplan-Meier estimator, which estimates the distribution function for censored data. As I had worked with Markov chains, I could immediately see that I could write the estimator as a product integral.

Of course this doesn't help anyone, but a product integral satisfies an obvious differential equation. And once you can express the estimator as the solution of a differential equation, you can find the asymptotic distribution, by doing the asymptotics on the equation instead of the solution. So we found the asymptotic distribution of what has later been called the Aalen-Johansen estimator, see Aalen and Johansen (1978).

## How did this come about?

The breakthrough in this area of Statistics came with the work of David Cox, who in 1972 presented the Cox survival model (Cox 1972) in which you model the hazard rate. That is, the intensity of the event under consideration, unemployment for instance, in a small interval around time $t$, given the past history. The hazard function is allowed to
depend on explanatory regressors. The expression for the likelihood then becomes a special case of the product integral.

In our department Niels Keiding worked with statistical methods applied to medical problems. He got interested in survival analysis and wanted to understand the mathematical theory behind it, so he was teaching the theory of point processes and martingales. A typical example of such problems is to follow a group of patients for a period to see, for instance, how a treatment is helping cure a disease. Ideally you follow all patients for as long as it takes, but in practice you have to terminate the study after a period, so the data is truncated.

The data is made more complicated to work with, because people can leave the study for other reasons, and hence the data is censored. Such data consists of a sequence of time points, and is therefore called a point process. Niels was very active with this type of data and he and his colleagues wrote the book Andersen et al. (1992), describing both applications and the theory of the analysis, including some of my work with product integration with Richard Gill.

## Did this research have practical implications?

At the University of Copenhagen a retrospective study of the painters syndrome was conducted. The reason for and time point of retirement were noted for a group of painters, and as control group the same data were recorded for bricklayers. Such data is typically made more complicated by individuals changing profession or moving, or dying during the period of investigation.

One way of analysing such data is to draw a plot of the estimated integrated intensity of retirement due to brain damage (painters' syndrome), which can take into account the censoring. It was obvious from that plot that the risk of brain damage was much higher for painters than for brick layers. This investigation was just a small part of a larger investigation which resulted in changing working conditions for painters, and much more emphasis on water based paint.

## Tell us about your work on convexity.

The topic was suggested to me by Hans Brøns shortly after I finished my studies and I had the opportunity to go to Berkeley for a year. The purpose was to write a thesis on the applications of convexity in probability. The important result in functional analysis was the theorem by Hewitt and Savage (Hewitt and Savage 1955) about representing points in a convex set as a mixture of extreme points. We hoped to find some applications of this result in probability theory.

The simplest example of such a result is that a triangle is a convex set with three extreme points, and putting some weights on the extreme points, we can balance the triangle by supporting it at its center of gravity, which is the weighted average of the extreme points. Another simple example is the set of Markov probability matrices, with positive entries adding to one in each row. The extreme points are of course the matrices you get by letting each row be a unit vector.

A more complicated example is the following: in probability theory there is a well known Lévy-Khintchine representation theorem, which says that the logarithm of the characteristic function of an infinitely divisible distribution is an integral of a suitable kernel with respect to a measure on the real line. It is not difficult to show that these functions form a compact convex set. One can identify the extreme points to be either Poisson distributions or the Gauss distribution. The representation theorem then follows from the result of Hewitt and Savage. This provided a new understanding and a new proof of the Lévy-Khintchine result.

Another result I worked on I still find very intriguing. If you consider a non-negative concave continuous function on the unit circle, normalized to have an integral of 1 , then such functions form a convex compact set. The challenge is to find the extreme points. I found a large class of extreme points, which have the property that they are piecewise flat.

I needed a further property: that at each corner of the function, where the flat pieces meet, there are only three pieces meeting. Imagine a pyramid with four sides, so that four lines meet at the top. This function is not an extreme point, but if you cut the tip off the pyramid, then at each of the four corners created will have only three sides meeting, and then it is an extreme point.

The set of functions has the strange property that each point in the set (a concave function) can be approximated uniformly close by just one extreme point.

## Tell us about other models you worked on.

I once collaborated with a group of doctors who were investigating the metabolism of sugar, say, by the liver in order to find a good measure of liver capacity. The data was the concentration of sugar in the blood at the inlet and the outlet of the liver. There were three models around at the time. One modelled the measurement at the inlet and the other at the outlet.

In developing the model we used an old idea of August Krogh-Winner of the Nobel Prize in Physiology or Medicine in 1920 "for his discovery of the capillary motor regulating mechanism"-of modelling the liver as a tube lined with liver cells on the inside, such that the concentration of sugar at the inlet would be higher than the concentration at the outlet. This physiological model gave the functional form of the relation between the inlet and outlet concentrations, which we used to model the data.

We used the data to compare the three models and found out that ours was the best. I worked on this with Susanne Keiding, see Keiding et al. (1979) and Johansen and Keiding (1981). We analyzed the data by nonlinear regression that used the mathematics of the model, the so called Michaelis-Menten kinetics.

## It is not so common to check model assumptions as suggested by David Hendry. What is your view on this?

In my own training in mathematics, I could not use a theorem without checking its assumptions. This is obviously in the nature of mathematics. Our education in Statistics was based on Mathematics, so for me it was natural to check assumptions when you have formulated a model for the data.

At the Economics Department of the University of Copenhagen Katarina held for 9 years a "Summer School in the Cointegrated VAR model: Methodology and Applications". In total we had about 300 participants. I would give the theoretical lectures and Katarina would tell them about how they need to model the data in order to investigate the economic theories.

The main aspect of the course, however, was that they brought their own data and had a specific economic question in mind concerning their favourite economic theory. They spent all afternoons for a month doing applied work, choosing and fitting a model, checking assumptions of the model, and comparing the outcome with economic knowledge they had. Katarina would supervise the students, and they were encouraged to discuss among themselves. They had never tried such a thing and learned a tremendous amount.

On a smaller scale, most courses should include some software for doing econometric analysis. Such programs would often produce output for different models (different lag length, cointegration rank, and deterministic terms) as well as misspecification tests. It seems a good idea to include the interpretation of such output in a course, so one can have a discussion of what it means for a model to be wrong, and how one can react to change it for the better.

## Is the ability to check assumptions related to likelihood models-i.e., models with a likelihood?

A very simple regression model, that everyone knows about, is to assume for two series $X_{t}, Y_{t}$ that they are linearly related $Y_{t}=\beta X_{t}+\varepsilon_{t}$, and the error terms $\varepsilon_{t}$ are mutually independent and independent of the $X_{t} s$. Obviously, without specifying a precise family of distributions for the error term, one cannot talk about likelihood methods. So what do we
gain by assuming Gaussian errors for example? We can derive the least squares method, but in fact Gauss did the opposite. He derived the distribution that gave you least squares.

There is another application of a parametric model that is also useful. Suppose you realize, somehow, that the regression residuals are autocorrelated. Then you would like to change the estimation method, and a method for doing that is to build a new model, which can tell you how to change the method. This is where an autoregressive model for $\varepsilon_{t}$ would, after a suitable analysis of the likelihood, the score, and information, tell you what to do.

So I think the answer is that the likelihood method tells you how to get on with the analysis, and what to do when your assumptions fail. In this light, one can see that the failure of conducting inference using a cointegrating regression can be remedied by formulating the CVAR with Gaussian errors and then derive the methods from the likelihood.

## How did your training help you, and what does this suggest for education needs in the econometric profession?

I think what helped me in Econometrics is the basic training I received in Mathematical Statistics. At the University of Copenhagen, the degree in Statistics ("candidatus statisticae") was introduced in 1960, when Anders Hjorth Hald was appointed professor of Statistics. He appointed Hans Brøns as the second teacher.

Anders Hald had been working for a number of years as statistical consultant and later as professor of Statistics at the Economic Department at the University of Copenhagen. He was inspired by the ideas of R. A. Fisher at Cambridge, and our Statistics courses were based on the concept of a statistical model and analysis of estimators and test statistics derived from the likelihood function. The purpose was to educate statisticians to do consulting with other scientists, but also to develop new statistical methods. The teaching was research-based and included many courses in mathematics.

The teaching attracted very good students. In those days, if you had a background in mathematics, there was essentially only one thing you could use it for, and that was teaching at high school. I was very interested in mathematics but did not want to teach at high school, so I became a statistician. This would allow me to collaborate with scientists from other fields, something that I would enjoy a lot.

Our department grew over the years to about 10 people and we discussed teaching and research full time. It was a very inspiring environment for exchanging ideas and results. We regularly had visitors from abroad, who stayed for a year doing teaching and research. For my later interest in Econometrics the course by Patrick Billingsley in 1964-1965 was extremely useful, as it taught me advanced probability theory. He was lecturing on what was to become the now classical book on convergence of probability measures (Billingsley 1968) while he was visiting Copenhagen.

## What should one do when the model doesn't fit?

There does not seem to be an easy set of rules for building models, so it is probably best to gain experience by working with examples. Obviously a model should be designed so it can be used for whatever purpose the data was collected. But if the first attempt fails, because it does not describe the data sufficiently well, it would possibly be a good idea to improve the model by taking into account in what sense it broke down.

You could look for more explanatory variables, including dummies for outliers, different variance structure or perhaps study related problems from other countries, say, to get ideas about what others do. It is my strong conviction that the parametric model can help you develop new estimation and test methods to help to find a model which better takes into account the variation of the data.

As students, we only analysed real life data and sometimes even had a small collaboration with the person who had taken the measurements. Our role would be to help building a statistical model and formulate the relevant hypotheses to be investigated in collaboration with the user. Then we would do the statistical analysis of the model based
on the likelihood function. With this type of training we learned to discuss and collaborate with others.

## How and why should models be built?

I do not think that there are general rules for model building, partly because models can serve so many different purposes. By considering many examples, it is my opinion that you can develop a feeling for what you do with the kind of problems you are investigating. But if you change field, you probably have to start from scratch. Thus the more experience you have with different types of models, the more likely it is that you can find a good model next time you need it.

I personally find that the main reason for building and analyzing models is that you want to be able to express your own understanding of the phenomenon to other people. The mathematical language has this nice property that you can communicate concepts in a precise way. I think about the model as a consistent way of formulating your understanding of the real world.

It is interesting to consider an average of measurements as something very relevant and useful in real life. The model for i.i.d. variables includes the nice result of the law of large numbers, and gives us a way of relating an average to an abstract concept of expectation in a model. But perhaps more important than that, is that it formulates assumptions, under which the result is valid, and that gives you a way of checking if the average is actually a good idea to calculate for the data at hand.

Another practically interesting concept is the notion of spurious correlation, which for nonstationary data can be very confusing, if you do not have a model as a basis for the discussion, see for instance the discussion of Yule (1926). It was the confusion about the notion of correlation (for nonstationary time series variables) that inspired the work of Clive on the concept of cointegration.

## Could you elaborate on the theory and practice of likelihood methods in Econometrics?

Econometric textbooks often contain likelihood methods, but they do not have a prominent position. There are only few books which are based on likelihood methods from the beginning, as for instance Hendry and Nielsen (2007). In the space between models and methods, the weight is usually on the methods and how they perform under various assumptions. There are two good reasons to read textbooks, one is that you can then apply the methods, and the other is that you can then design new methods.

When R. A. Fisher introduced likelihood analysis, the starting point was obviously the model, and the idea is that the method for analysing the data should be derived from the model. In fact it is a unifying framework for deriving methods that people would be using anyway. Thus instead of remembering many estimators and statistics, you just need to know one principle, but of course at the price of some mathematical analysis.

By deriving the method from first principles you also become more aware of the conditions for the analysis to hold, and that helps checking for model misspecification, which again can help you modify the model if it needs improvement. It is clear that the likelihood requires a model, and the likelihood analysis is a general principle for deriving estimators and test statistics; yet it usually also requires a lot of mathematical analysis, and the solutions often need complicated calculations.

It is, however, not a solution to all problems, there are counter examples. In particular, when the number of parameters increases, the maximum likelihood estimator can be inconsistent. A standard example is to consider observations ( $X_{i}, Y_{i}$ ), $i=1, \ldots, n$ which are independent Gaussian with mean $\left(\mu_{i}, \mu_{i}\right)$ and variance $\sigma^{2}$. In this simple situation $\widehat{\sigma}^{2} \rightarrow \frac{1}{2} \sigma^{2}$ in probability, the so-called "Neyman-Scott Paradox".

## What are the alternative approaches with respect to a well-specified statistical model?

The simple regression model, where the calculations needed to find the estimator is the starting point, is an example of an algorithm which is often taken as the starting point, and which does not require a statistical model. The statistical model is needed, when you want to test hypotheses on the coefficients, and the parametric statistical model is useful if you want to derive new methods.

Of course there exists many methods, expert systems, based on complicated nonlinear regressions. I am not an expert on these, but I note that the people behind them collaborate with statisticians.

## So what needs to be avoided is the use of Statistics without knowledge of it. Correct?

Sounds like a good idea! Many people think that Statistics is a set of well developed methods that we can just use. I think that can be a bit dangerous, and highly unsatisfactory for the users. It would of course be lovely, but a bit unrealistic, that all users should have a deep understanding of Statistics, before they could use a statistical method. I explained elsewhere the summer course we had in Copenhagen, where the students are put in a situation where they have to make up their minds about what to do, and that certainly improves learning.

## Are statisticians especially trained to collaborate?

As a statistician, you study all the classical models about Poisson regression and two-way analysis of variance, survival analysis and many more. If the exercises contain real data, you will learn to formulate and build models and choose the right methods for analyzing them. It is of course in the nature of the topic that if you are employed later in a medical company doing controlled clinical trials, then you will have to collaborate with the doctors.

The education should therefore also try to put the students in situations, where such skills can be learned. The problem is of course that if you end up in an insurance company or in an economics department you probably need different specializations. So in short I think the answer to your question is Yes! the students should be trained to collaborate.

## Hence, is Statistics a science at the service of other sciences?

Of course Statistics as a field has a lot of researchers working at Universities on teaching and developing the field, but most statisticians work in industry or public offices, pharmaceutical companies, insurance companies, or banks.

Another way of thinking about it was implemented by my colleague Niels Keiding. In 1978 he started a consulting service for the medical profession at the University of Copenhagen, using a grant from the Research Council. The idea was to help the university staff in the medical field getting expert help with their statistical problems, from planning controlled clinical trials, to analysing data of various sort. This has been a tremendous success and now is a department at the University with around 20 people working full time in this, as well as some teaching of Statistics for the doctors.

## Any message on the publication process?

I remember when I was in Berkeley many years ago, in 1965, I took a course with Lester Dubins, who had just written a book called "How to gamble if you must", Dubins and Savage (1965). I was then working with the coauthor on my first paper, Johansen and Karush (1966). I must have been discussing publications with Lester, and he kindly told me "But you have to remember, Søren: every time you write a paper and get it published, it becomes slightly more difficult for everybody else to find what they want".

This carried a dual message on the benefit of advancing knowledge and the associated increased cost in retrieving information. Fortunately, this cost has been greatly reduced by the current powerful internet search engines available.

Author Contributions: R.M. and P.P. have contributed to all phases of the work. All authors have read and agreed to the published version of the manuscript.
Funding: This research received no external funding.
Institutional Review Board Statement: Not applicablet.
Informed Consent Statement: Not applicable.
Data Availability Statement: Not applicable.
Acknowledgments: Authors gratefully acknowledge useful comments and correction from the interviewee on the draft of the article.
Conflicts of Interest: The authors declare no conflict of interest. Information and views set out in this paper are those of the authors and do not necessarily reflect the ones of the institutions of affiliation.

## Notes

Accessed on 3 April 2022.
2 See Engle and Granger (1987) and Johansen (1988b).
3 For a description and the current version of PcGive see: https://www.doornik.com/doc/PcGive/, accessed on 3 April 2022.

## References

Aalen, Odd O., and Søren Johansen. 1978. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. Scandinavian Journal of Statistics 5: 141-50.
Ahn, Sung K., and Gregory C. Reinsel. 1990. Estimation for partially nonstationary multivariate autoregressive models. Journal of the American Statistical Association 85: 813-23. [CrossRef]
Andersen, Per K., Ørnulf Borgan, Richard D. Gill, and Niels Keiding. 1992. Statistical Models Based on Counting Processes. New York: Springer.
Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Heiko Ebens. 2001. The distribution of realized stock return volatility. Journal of Financial Economics 61: 43-76. [CrossRef]
Anderson, Theodore Wilbur. 1951. Estimating linear restrictions on regression coefficients for multivariate normal distributions. The Annals of Mathematical Statistics 22: 327-51. [CrossRef]
Billingsley, Patrick. 1968. Convergence of Probability Measures. New York: John Wiley \& Sons.
Campbell, John Y., and Robert J. Shiller. 1987. Cointegration and tests of present value models. Journal of Political Economy 95: 1062-88. [CrossRef]
Cox, David R. 1972. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) 34: 187-202.
Davidson, James E. H., David F. Hendry, Frank Srba, and Stephen Yeo. 1978. Econometric Modelling of the Aggregate Time-Series Relationship Between Consumers' Expenditure and Income in the United Kingdom. The Economic Journal 88: 661-92. [CrossRef]
Diebold, Francis X., and Atsushi Inoue. 2001. Long memory and regime switching. Journal of Econometrics 105: 131-59.
Dubins, Lester E., and Leonard Savage. 1965. How to Gamble if You Must: Inequalities for Stochastic Processes. New York: McGraw-Hill.
Engle, Robert F., and Byung Sam Yoo. 1991. Cointegrated time series: An overview with new results. In Long Run Economic Relationships: Readings in Cointegration. Edited by Robert F. Engle and Clive W. J. Granger. Oxford: Oxford University Press, chp. 12, pp. 237-66.
Engle, Robert F., and Clive W. J. Granger. 1987. Co-integration and Error Correction: Representation, Estimation, and Testing. Econometrica 55: 251-76. [CrossRef]
Fisher, Franklin M. 1966. The Identification Problem in Econometrics. New York: McGraw-Hill.
Gill, Richard D., and Søren Johansen. 1990. A Survey of Product-Integration with a View Toward Application in Survival Analysis. The Annals of Statistics 18: 1501-55. [CrossRef]
Granger, Clive W. J. 1980. Long memory relationships and the aggregation of dynamic models. Journal of Econometrics 14: 227-38. [CrossRef]
Hall, Anthony D., Heather M. Anderson, and Clive W. J. Granger. 1992. A cointegration analysis of treasury bill yields. The Review of Economics and Statistics 74: 116-126. [CrossRef]
Hansen, Bruce E. 2018. Johansen's reduced rank estimator is GMM. Econometrics 6: 26. [CrossRef]
Hendry, David F., and Bent Nielsen. 2007. Econometric Modeling: A Likelihood Approach. Princeton: Princeton University Press.
Hewitt, Edwin, and Leonard J. Savage. 1955. Symmetric measures on cartesian products. Transactions of the American Mathematical Society 80: 470-501. [CrossRef]
Hillebrand, Eric, Søren Johansen, and Torben Schmith. 2020. Data revisions and the statistical relation of global mean sea level and surface temperature. Econometrics 8: 41. [CrossRef]
Howlett, Phil G. 1982. Input retrieval in finite dimensional linear systems. Journal of Australian Mathematical Society (Series B) 23 : 357-82. [CrossRef]
Johansen, Søren. 1988a. The mathematical structure of error correction models. Contemporary Mathematics 80: 359-86.

Johansen, Søren. 1988b. Statistical analysis of cointegration vectors. Journal of Economic Dynamics and Control 12: 231-54. [CrossRef] Johansen, Søren. 1992. A representation of vector autoregressive processes integrated of order 2. Econometric Theory 8: 188-202. [CrossRef]
Johansen, Søren. 1995a. Identifying restrictions of linear equations with applications to simultaneous equations and cointegration. Journal of Econometrics 69: 111-132.
Johansen, Søren. 1995b. A statistical analysis of cointegration for I(2) variables. Econometric Theory 11: 25-59. [CrossRef]
Johansen, Søren. 1996. Likelihood-Based Inference in Cointegrated Vector Autoregressive Models, 2nd ed. Oxford: Oxford University.
Johansen, Søren. 1997. Likelihood analysis of the I(2) model. Scandinavian Journal of Statistics 24: 433-62. [CrossRef]
Johansen, Søren. 2008. A representation theory for a class of vector autoregressive models for fractional processes. Econometric Theory 24: 651-76. [CrossRef]
Johansen, Søren. 2019. Cointegration and adjustment in the $\operatorname{CVAR}(\infty)$ representation of some partially observed $\operatorname{CVAR}(1)$ models. Econometrics 7: 2. [CrossRef]
Johansen, Søren, and Bent Nielsen. 2009. An analysis of the indicator saturation estimator as a robust regression estimator. In The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry. Edited by Jennifer Castle and Neil Shephard. Oxford: Oxford University Press, pp. 1-36.
Johansen, Søren, and Bent Nielsen. 2016a. Analysis of the forward search using some new results for martingales and empirical processes. Bernoulli 22: 1131-83. [CrossRef]
Johansen, Søren, and Bent Nielsen. 2016b. Asymptotic theory of outlier detection algorithms for linear time series regression models. Scandinavian Journal of Statistics 43: 321-48. [CrossRef]
Johansen, Søren, and Bent Nielsen. 2019. Boundedness of M-estimators for linear regression in time series. Econometric Theory 35: 653-83. [CrossRef]
Johansen, Søren, and Jack Karush. 1966. On the semimartingale convergence theorem. The Annals of Mathematical Statistics 37: 690-94. [CrossRef]
Johansen, Søren, and Katarina Juselius. 1990. Maximum likelihood estimation and inference on cointegration-With applications to the demand for money. Oxford Bulletin of Economics and Statistics 52: 169-210. [CrossRef]
Johansen, Søren, and Morten Ørregaard Nielsen. 2010. Likelihood inference for a nonstationary fractional autoregressive model. Journal of Econometrics 158: 51-66.
Johansen, Søren, and Morten Ørregaard Nielsen. 2012. Likelihood inference for a fractionally cointegrated vector autoregressive model. Econometrica 80: 2667-732. [CrossRef]
Johansen, Søren, and Morten Ørregaard Nielsen. 2018. The cointegrated vector autoregressive model with general deterministic terms. Journal of Econometrics 202: 214-29. [CrossRef]
Johansen, Søren, and Fred L. Ramsey. 1979. A bang-bang representation for $3 \times 3$ embeddable stochastic matrices. Zeitschrift fur Wahrschlichkeitsteorie und Verwandte Gebiete 47: 107-18. [CrossRef]
Johansen, Søren, and Susanne Keiding. 1981. A family of models for the elimination of substrate in the liver. Journal of Theoretical Biology 89: 549-56. [CrossRef]
Johansen, Søren, Henrik Hansen, and Stefano Fachin. 2005. A Simulation Study of some Functionals of Random Walk. Technical Report, University of Copenhagen. Available online: http:/ /web.math.ku.dk/ ~sjo / (accessed on 3 April 2022).
Juselius, Katarina, ed. 1994. Econometric Modelling of Long-run Relations and Common Trends: Theory and Applications. Copenhagen: University of Copenhagen, Institute of Economics, vols. I-IV.
Keiding, Susanne, Søren Johansen, Inge Midtbøll, Alan Rabøl, and Leif Christiansen. 1979. Ethanol elimination kinetics in human liver and pig liver in vivo. American Journal of Physiology-Endocrinology and Metabolism 237: E316. [CrossRef]
Nielsen, Heino Bohn, and Anders Rahbek. 2007. The likelihood ratio test for cointegration ranks in the I(2) model. Econometric Theory 23: 615-37. [CrossRef]
Onatski, Alexei, and Chen Wang. 2018. Alternative asymptotics for cointegration tests in large VARs. Econometrica 86: 1465-78. [CrossRef]
Parke, William R. 1999. What is Fractional Integration? The Review of Economics and Statistics 81: 632-38. [CrossRef]
Phillips, Peter C. B. 1986. Understanding spurious regressions in econometrics. Journal of Econometrics 33: 311-40. [CrossRef]
Phillips, Peter C. B. 1995. Fully modified least squares and vector autoregression. Econometrica 63: 1023-78. [CrossRef]
Santos, Carlos, David F. Hendry, and Søren Johansen. 2008. Automatic selection of indicators in a fully saturated regression. Computational Statistics 23: 317-35. correction 337-39. [CrossRef]
Velu, Raja P., Gregory C. Reinsel, and Dean W. Wichern. 1986. Reduced rank models for multiple time series. Biometrika 73: 105-18. [CrossRef]
Welsh, Dominic J. A. 1976. Matroid Theory. London: Academic Press.
Yule, G. Udny. 1926. Why do we sometimes get nonsense-correlations between time-series?-A study in sampling and the nature of time-series. Journal of the Royal Statistical Society 89: 1-63. [CrossRef]

## Article

# Searching for a Theory That Fits the Data: A Personal Research Odyssey 

Katarina Juselius

Citation: Juselius, Katarina. 2021. Searching for a Theory That Fits the Data: A Personal Research Odyssey. Econometrics 9: 5. https://doi.org/ 10.3390/econometrics 9010005

Received: 6 July 2018
Accepted: 11 January 2021
Published: 1 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations


Copyright: © 2021 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/).

Department of Economics, University of Copenhagen, DK-1353 Copenhagen, Denmark; katarina.juselius@econ.ku.dk


#### Abstract

This survey paper discusses the Cointegrated Vector AutoRegressive (CVAR) methodology and how it has evolved over the past 30 years. It describes major steps in the econometric development, discusses problems to be solved when confronting theory with the data, and, as a solution, proposes a so-called theory-consistent CVAR scenario. A number of early CVAR applications are motivated by the urge to find out why the empirical results did not support Milton Friedman's concept of monetary inflation. The paper also proposes a method for combining partial CVAR analyses into a large-scale macroeconomic model. It argues that an empirically-based approach to macroeconomics preferably should be based on Keynesian disequilibrium economics, where imperfect knowledge expectations replace so called rational expectations and where the financial sector plays a key role for understanding the long persistent movements in the data. Finally, the paper argues that the CVAR is potentially a candidate for Haavelmo's "design of experiment for passive observations" and provides several illustrations.


Keywords: cointegrated VAR; methodology; linking theory to evidence; empirically-based macroeconomics

JEL Classification: B41; C32; C51; C52

## 1. Introduction

I was happy to accept the invitation by the guest editors to write this survey paper based on my retirement lecture given at the Economics Department of the University of Copenhagen in 2014. Retirement is one of the important dividing lines in a long active life that gives you the opportunity to slow down and to reflect on your achievements. When preparing for my retirement lecture I asked myself: who inspired me to choose econometrics; what were the main questions that motivated my research; how did I go about answering them; what stones did I stumbled on; and the most important one: did my research contribute to useful answers of important questions. In writing this paper, I have allowed myself to focus almost exclusively on my own research together with my many coauthors. While the paper is far from a balanced account of all the good research that has inspired me, the bibliographies in the papers to be discussed bear witness of the many important contributions on which this research rests.

Over a long academic career it is almost unavoidable that some scholars have been more influential than others. For me, David Hendry and Clive Granger were enormously influential for my thinking in the early formative years. I found the "general-to-specific" error correction approach developed by David utterly exciting and the numerous timeseries methods proposed by Clive very inspiring. To be a colleague and a friend of both of them has been an invaluable privilege in all these years. ${ }^{1}$ My research has benefitted a lot from their highly innovative research.

[^0]However, it was the working paper on cointegration and error correction (Granger 1983) that fundamentally changed both my professional career and my personal life. From the outset I was intrigued by the concept of cointegration and how it related to the more familiar concept of error-correction. Clive's paper defined cointegration as part of a vector moving average model for unobservable errors, whereas error correction models were based on the autoregressive model formulated for variables. At that time it was difficult to estimate moving average models-definitely more so than error correction models-and I could not see how to use cointegration in empirical work. Therefore, I asked Søren Johansen to give a prepared comment on Clive's paper at the Nordic Statisticians meeting in 1982. Søren, recognizing the great potential of Clive's cointegration idea as a means for solving the problem of nonstationarity in economic time-series processes, gave an insightful presentation.

As most economic time series are nonstationary, but the statistical theory used to analyze them was based on the assumption of stationarity, this was clearly extremely important. One can say that we stumbled over a gold mine of relevant problems that needed to be solved. The first one was to formulate the concept of cointegration in the context of a vector autoregressive model. With Søren's formal training in mathematical statistics it did not take long until he had derived a rigorous solution in terms of an autoregression with a reduced rank impact matrix, as well as a maximum likelihood solution for its estimation based on reduced rank regression. Many more useful results followed in a steady stream. I was thrilled-and still am—by the numerous possibilities that cointegration analysis offers to ask new and relevant questions in economics.

In the mid-nineties, most of the econometric tools needed for a full-fledged Cointegrated Vector AutoRegressive (CVAR) analysis were derived and I could start focusing on what interested me most: to develop the CVAR as an empirical methodology in macroeconomics. I had come across Trygve Haavelmo's Nobel Prize winning monograph "The Probability Approach to Economics" (Haavelmo 1944) and was immediately struck by its beauty. Trygve Haavelmo, as it appeared, had already-before I was born-formulated a stringent vision of a likelihood based approach to economic modeling that seemed to be the answer to my own rather muddled methodological questions. Haavelmo's concept of a "designed experiment for data by passive observations" was exactly what I needed when I struggled to work out how to associate the theoretical structures of macroeconomic models with the much richer structures of the CVAR model.

Common to almost all my empirical papers was the puzzlement that the CVAR results in one way or the other seemed to contradict basic assumptions of the underlying economic theory. Especially in the early years, it was something I was strongly worried about: had I misunderstood something crucial? Did I apply the CVAR in the correct way? I happened to stumble over a methodology book by David Colander and then read almost everything I could find from his pen. His thorough insight in the methodology of economics helped me see that the problems were not necessarily related to the CVAR model.

All this and much more is discussed in the rest of the paper which is organized around four major themes.

The first one is about the development of the econometric foundations of the CVAR and describes (i) major stepping stones that were needed in order to apply cointegration techniques to relevant economic problems, (ii) my first attempts to confront economic theories with data and my puzzlement when results did not support standard economic assumptions, and (iii) the development of a user-friendly software.

The second theme is about the development of the CVAR as an empirical methodology and describes (i) numerous difficulties to be solved when confronting economic theories with the data; (ii) my many efforts to formulate a viable link between the economic model and the data as structured by the CVAR, which finally lead to the concept of a so called theory-consistent CVAR scenario; and (iii) my attempts to associate the CVAR approach with Trygve Haavelmo's probability approach to economics.

The third theme is about early applications starting with the Danish money demand which was primarily used as a check of the derived econometric results but also to understand the mechanisms governing price inflation. The Danish money demand study is about successes but also puzzling results, which forced me to search for alternative explanations to inflation pressure. Finally, this part discusses a procedure for how to combine partial CVAR models into a larger model in which all aspects of the inflationary mechanism can be studied.

The forth theme is about a new approach to empirical macro. The long persistent swings in the data are tentatively explained by replacing rational expectations with imperfect knowledge expectations. In particular, real exchange persistence is related to speculative behavior in foreign currency markets affecting nominal exchange rates but not consumer prices. This part also discusses why persistent long swings in real asset prices are prone to generate long swings in the real economy, particularly in the unemployment rate. The potential of the CVAR to act as a "design of experiment" in macroeconomics is illustrated with unemployment dynamics in a crisis period based on the Finnish house price crisis in the nineties and the recent Greek depression.

The paper ends with some personal reflections on obstacles and bumps on the long journey and concludes with a discussion of what we should require from empirically relevant macroeconomics.

## 2. Econometric Foundations

The starting point of the cointegration project was the unrestricted $\operatorname{VAR}(k)$ model:

$$
\begin{align*}
\Delta x_{t} & =\Pi x_{t-1}+\sum_{i=1}^{k-1} \Gamma_{i} \Delta x_{t-i}+\mu_{0}+\mu_{1} t+\Phi_{1} D_{t}+\Phi_{2} S_{t}+\varepsilon_{t}  \tag{1}\\
t & =1, \ldots, T
\end{align*}
$$

where $x_{t}$ is a $p \times 1$ data vector, $\mu_{0}$ is a $p \times 1$ vector of constant terms $\mu_{1}$ a $p \times 1$ vector of trend coefficients, $D_{t}$ a a $m \times 1$ vector of dummy variables, $S_{t}$ an $s \times 1$ vector of seasonal dummies, and $\varepsilon_{t} \sim \operatorname{Niid}(0, \Omega)$. In the first years, (1) was analyzed without a linear trend and dummies in the model. However, as most macroeconomic data are trending and riddled with extreme events, it did not take long before we realized that both are indispensable for an adequately specified model. This led Johansen (1994) to discuss the dual role of the constant and the trend in the CVAR and to provide a solution. With time we learned the simple lesson that the choice of VAR specification from the outset should be either for non-trending data with $\mu_{1}=0$ and with the constant restricted to the cointegration relations, or for trending data with $\mu_{1} \neq 0$ and with the trend restricted to the cointegration relations.

When $x_{t}$ is integrated of order one, $I(1)$, all components in (1) except $\Pi x_{t-1}$ are stationary. Therefore, either $\Pi=0$ or of reduced rank, $r$. It was a defining moment when Søren in 1986 was able to find the likelihood-based solution to $\Pi=\alpha \beta^{\prime}$ where $\alpha, \beta$ are $p \times r$. After that it was possible to address economic problems in an $I(1)$ world using a likelihood based VAR analysis.

Inverting (1) with $\Pi=\alpha \beta^{\prime}$ allowed us to express the vector, $x_{t}$, as a function of the shocks, $\varepsilon_{t}$, and the deterministic terms constant, trend, and dummies:

$$
\begin{equation*}
x_{t}=\beta_{\perp}\left(\alpha_{\perp}^{\prime} \Gamma \beta_{\perp}\right)^{-1} \alpha_{\perp}^{\prime} \sum_{i=1}^{t}\left(\varepsilon_{i}+\Phi_{1} D_{i}+\mu_{0}+\mu_{1} i\right)+C^{*}(L)\left(\Phi_{1} D_{t}+\mu_{0}+\mu_{1} t+\varepsilon_{t}\right)+A \tag{2}
\end{equation*}
$$

where $\Gamma=I-\sum_{i=1}^{k-1} \Gamma_{i}, A$ depends on initial values and $\beta^{\prime} A=0, \beta_{\perp}$ and $\alpha_{\perp}$ are $p \times$ $p-r$ matrices orthogonal to $\beta$ and $\alpha, \alpha_{\perp}^{\prime} \sum_{i=1}^{t} \varepsilon_{i}$, is a measure of the $p-r$ stochastic trends, $\beta_{\perp}\left(\alpha_{\perp}^{\prime} \Gamma \beta_{\perp}\right)^{-1}$ denote the coefficients with which the stochastic trends load into the variables, and $C^{*}(L)\left(\Phi_{1} D_{t}+\mu_{0}+\mu_{1} t+\varepsilon_{t}\right)$ represents stationary movements around
the trends. The formulation (2) allowed us to calculate impulse response functions and long-run dynamical effects of exogenous shocks to the system, the so called long-run multiplier effects.

Economic data frequently exhibit too much persistence to be tenable with the $I(1)$ assumption. The condition that $x_{t}$ is $I(2)$, i.e., nonstationary of second order, was formulated in Johansen (1992) as the reduced rank of $\alpha_{\perp} \Gamma \beta_{\perp}^{\prime}=\xi \eta^{\prime}$, where $\xi, \eta$ are $p-r \times s$.

Fortunately, it was only the cointegration rank test that needed a nonstandard distribution. After the rank was found, the nonstationary data was transformed to stationarity partly by differencing and partly by taking stationary linear combinations of the levels: this leads to standard Gaussian and $\chi^{2}$ asymptotic inference in the transformed model. A large number of important economic hypotheses, such as exogeneity, endogeneity, long-run homogeneity, identifying restrictions, zero restrictions, etc. could then be tested using standard procedures.

### 2.1. Econometric Theory and Economic Applications

Already in 1986, Søren worked out the representation theory, the probability theory, and the statistical theory that were necessary for applying likelihood-based cointegration analysis to empirical problems. The results were subsequently published in Johansen (1988). At the same time as Søren derived the theoretical results, I applied them to the Danish data consisting of real money holdings (M2), real aggregate demand, a weighted deposit rate for M2, and the long-term bond rate. While the primary goal was to have a test case for the theoretical results, the ultimate goal was to obtain a likelihood based estimate of the Danish money demand relation for M2. Luckily, this relation turned out to be incredibly stable over time-possibly the most stable macroeconomic relation I have ever come across. This was invaluable as we were able to develop the main cointegration tools and test them based on data that gave reasonably interpretable results. Later on, we had ample possibilities to tackle more challenging problems which often forced us to rethink both econometrics and economics.

The results of this first "going back and forth" between econometric theory and money demand became a working paper in 1987. It was submitted to Econometrica, where it was lying for more than two years and then rejected. In 1990, it was finally published in Oxford Bulletin of Economics and Statistics and became highly cited. ${ }^{2}$ The paper discusses, theoretically and empirically, how to test and impose a reduced rank on the VAR model, how to test hypotheses on the cointegration parameters $\beta$, and on the adjustment coefficients $\alpha$. The trace test showed that the rank was one, which was fortunate as it greatly simplified the statistical analysis. It was also fortunate that the cointegration relation was readily interpretable as a deviation from a plausible long-run money demand relation. Furthermore, it turned out that money stock alone was adjusting to $\beta_{1}^{\prime} x_{t}$ with a significant $\alpha$ coefficient, i.e., all the remaining $\alpha$ coefficients could be set to zero. Johansen (1992) subsequently showed that this was the condition for when the CVAR estimates of the cointegration relation are equivalent to the ones obtained from a single equation error correction model.

In many ways, it was a rich paper illustrating a variety of the rather complex cointegration methods with a realistic application to macroeconomic data. It received a lot of interest both among econometricians and empirical macroeconomists and it therefore bothers me that the deterministic terms were not satisfactorily specified. Today, I would approach the empirical analysis somewhat differently.

The next joint paper, Johansen and Juselius (1992), discusses some additional tests on the cointegration relations $\beta^{\prime} x_{t}$ based on an empirical application to the purchasing power parity (PPP) and the uncovered interest rate parity (UIP) for UK data. The paper shows theoretically and empirically how to test the same restriction on all $\beta$ vectors, which corresponds to a transformation of the data vector, and how to test the stationarity of a

[^1]known vector in $\beta$, for example, the stationarity of the real interest rate. The latter test procedure was extended to the case where some of the coefficients of a cointegration vector are known but others have to be estimated, for example, the stationarity of the real interest rate with an equilibrium mean shift.

It was also the first application where some of the cointegration relations ( $\beta^{\prime} x_{t}$ ) looked nonstationary, but the same cointegration relations corrected for the short-run dynamics ( $\beta^{\prime} R_{t}$ ) seemed perfectly stationary. This puzzling feature led to the development of the $I(2)$ model as will be described below.

A third joint paper, Johansen and Juselius (1994), discusses the important issue of identification of the long-run cointegration structure in terms of formal, empirical, and economic identification. Formal identification is needed to ensure that the parameters are estimable, empirical identification that all parameters necessary for formal identification are statistically significant, and economic identification that the results make economic sense. The paper shows theoretically and empirically how to impose and test identifying restrictions on a full $\beta$ structure and discusses all three aspects of identification based on an IS-LM model for Australian data.

With these three papers, the basic tools for a realistic analysis of economic problems in a nonstationary $I(1)$ world had been worked out. This was sufficient as long as economic data were assumed to be either stationary or at most $I(1)$. However, as the puzzling empirical results in Johansen and Juselius (1992) showed, the possibility of $I(2)$ variables had to be taken seriously. The test of the $I(2)$ hypothesis was formally derived in Johansen (1992) and illustrated with an analysis of PPP and UIP between Australia and the USA. Juselius (1995) reported a similar analysis between Germany and Denmark. Common for these papers was the finding that at least one of the cointegration relations, $\beta^{\prime} x_{t}$, was nonstationary, whereas $\beta^{\prime} R_{t}$ (for which the short-run effects had been concentrated out) was definitely stationary. This made sense in a CVAR model where $x_{t} \sim I(2), \beta^{\prime} x_{t} \sim I(1)$, and $\left(\beta^{\prime} x_{t}+\omega^{\prime} \Delta x_{t}\right) \sim I(0)$, where $\omega^{\prime} \Delta x_{t} \sim I(1)$. Based on the so-called two-step procedure, it was then straightforward to estimate and analyze the $I(2)$ model. Subsequently the two-step procedure was replaced by the likelihood based procedure in Johansen (1997). Juselius (1999a) used the likelihood based procedure to study long-run and medium-run price homogeneity among six US price indices.

Thus, the $I(2)$ analysis was initiated by trying to understand why the empirical results looked so strange, illustrating that the theoretical advances often were motivated by empirical necessity.

In the mid-nineties, most of the CVAR theory was developed and all ingredients needed for a successful cointegration analysis were available. Cointegration had become the standard way of analyzing economic time-series. The mathematical results needed for the probability/statistical analysis of cointegration were summarized by Søren in his book "Likelihood based inference in Cointegrated Vector Autoregressive Models" Johansen (1996). Ten years later, my book "The Cointegrated VAR model: Methodology and Applications" (Juselius 2006) was published, offering detailed discussions of the CVAR as an empirical methodology for macroeconomic applications.

In 1999, the Energy Journal commissioned David Hendry and myself to produce two expository papers on unit roots and cointegration for the readers of the journal. Hendry and Juselius (2000) explained the concepts in the context of a single equation error correction model, and Hendry and Juselius (2001) in the context of a system CVAR model. The two papers became highly cited also outside the field of energy economics demonstrating the profession's interest in applying cointegration in various branches of economics.

The appealing novelty of the CVAR model was that it was tailor-made to study longrun, medium-run, and short-run structures in the same model, allowing the complexity of the empirical reality to be grasped and better understood. Cointegration and the adjustment dynamics, the so-called pulling forces, were analyzed in the autoregressive representation of the model, while common trends, long-run multipliers and impulse response functions, the so-called pushing forces, were analyzed in the moving average representation. The

CVAR offered a detailed and immensely rich analysis of a variety of economic issues, including estimates of dynamic long-run effects of policy changes which had previously been difficult to estimate. Hoover et al. (2009) argue that "the CVAR model has a good chance of nesting a multivariate, path-dependent data-generating process and relevant dynamic macroeconomic theories". I was convinced that this approach would mean a big step forward toward an improved understanding of our macroeconomy.

### 2.2. Developing a User Friendly Software

Henrik Hansen translated our various program codes into a nice menu-driven package, CATS in RATS, version 1 (Hansen et al. 1994). It was the first software package to contain all the various tests and tools and the demand for it was correspondingly huge. However, the CVAR methodology was subject to an intense development and the need for an updated version grew for each year. In particular, we desperately needed a menu-driven program for a full-fledged $I(2)$ analysis based on likelihood-based principles. For two years, Jonathan Dennis worked extremely hard to produce the next version CATS in RATS, Version 2.0. (Dennis et al. 2006). It contained not just a full $I(2)$ analysis, but also a variety of new and improved features. Among others it added an expert system for long-run identification that greatly facilitated the search for empirically meaningful long-run structures in the data. It increased my own productivity enormously, probably by a factor of 50 or more. Recently, Jurgen Doornik translated the RATS code into OxMetrics and invested a huge amount of time and effort into the project. In particular the coding of the $I(2)$ analysis into OxMetrics was a major achievement. CATS, version 3.0 is now available (Doornik and Juselius (2017)).

## 3. The CVAR as an Empirical Methodology

From the outset, the idea of the CVAR was to offer a framework in which data would be allowed to speak freely without being silenced by prior restriction and in which basic hypotheses could be adequately tested and empirically relevant structures estimated. It is Popperian in this sense that the fundamental principle builds on the ability to falsify a hypothesis, to let the statistical analysis guide you toward an empirically relevant model. If the latter is inconsistent with your prior, then the analysis will often help you to see why your prior was wrong.

I was convinced this would make it possible to properly test the basic underlying assumptions of macroeconomic models and hoped it would replace the standard procedure of forcing the chosen theory model onto the data-also when they protest strongly. To my disappointment, not many economists seemed interested in having their models robustified or falsified in this fashion.

### 3.1. Confronting Theories with Data

While I never expected the empirical results to perfectly support standard theory, it came as a surprise that the results and conclusions differed so much. Discovering that some very fundamental relationships which most macroeconomic models relied on were not supported by the data was very disturbing and forced me to start thinking about methodological issues. After many unsuccessful attempts to interpret the CVAR results in terms of standard theory, it dawned on me that many economic theories might make more sense in a stationary than a non-stationary world. Few economic models at that time made an explicit distinction between stationary and nonstationary processes. Therefore, the idea of stochastic trends as the exogenous drivers of a system and dynamic adjustment to long-run equilibrium relations seemed foreign to most economists. Exogeneity played an important role but was differently defined in economics and econometrics. In the former case, it was essentially assumed, in the latter defined as weak, strong, and super exogeneity. The latter were formulated in terms of the statistical model and, thus, testable. See Engle et al. (1983).

Ever since the seminal paper by Sargan (1964), error correction models had been developed in numerous papers mostly by David Hendry and his followers. These were
mostly applied as single equation models and the error correction mechanism was assumed to be a measure of an equilibrium error. However, even these relatively simple and economically intuitive error correction models did not seem to exert much influence on standard economic thinking. What seemed to be needed, I thought, was a bridging principle that would link theoretical macroeconomic models in economics to the pulling and pushing forces of the CVAR model. Juselius (1993) was my first attempt to discuss this dichotomy in terms of a monetary problem without yet offering a bridging principle.

The ceteris paribus assumption-everything else constant or, more realistically, "everything else stationary"-was another issue I was concerned about. In a theoretical model, this assumption allows you to keep certain variables fixed and, therefore, to focus on those of specific interest. In an empirical model you have to bring these ceteris paribus variables into the analysis by conditioning. If they are stationary, the conclusions from the theoretical model are more likely to be robust, but if they are non-stationary, the conclusions can-and often do-change fundamentally. Because of this, it worried me that I frequently found important economic determinants such as the real interest rate, the real exchange rate, and the term spread to be empirically indistinguishable from a unit root process. In those cases when they are not explicitly part of the macroeconomic model, they are nonetheless part of the ceteris paribus clause. When these variables were included in the CVAR system I often found that conclusions changed, sometimes fundamentally so. The theory division of variables into endogenous, exogenous and fixed could not a priori be assumed to hold in the empirical model.

Expectations, which play such a prominent role in economic models, were problematic for CVAR models formulated in terms of observed variables. Economists usually solve this problem by making assumptions on how (rational) economic agents would forecast future outcomes given the chosen theoretical model-the so-called model based rational expectations hypothesis (REH). From the outset, I was skeptical of using REH as an empirical modeling device, mostly because I considered REH behavior to be highly unrealistic or even irrational in a nonstationary world with frequent breaks. The fact that Johansen and Swensen $(1999,2004)$ found essentially no support for the REH hypotheses when tested in the context of a CVAR model, only confirmed my doubts. Unfortunately, I had no clue how to solve the problem of unobserved expectations in a CVAR and for many years it was a constant worry. After stumbling over the theory of imperfect knowledge expectations, I began to see a possible way forward. However, it took me many attempts and a long time and until I was able to formulate a CVAR scenario that also included testable assumptions on theory-consistent expectations. See Juselius $(2017,2021 a)$.

Finally, there was the important issue of aggregation from the micro to the macro level. Most theoretical models in macroeconomics were based on the assumption of a representative agent. This simplifying assumption facilitated a mathematical formulation of the economic problem but often at the expense of its empirical relevance. It certainly seemed to be one reason why my empirical CVAR results deviated so strongly from the ones assumed in mainstream macroeconomic models.

The adoption of the Euro increased the interest in Euro-wide analyses and, therefore, the need to create sufficiently long historical data series aggregated over individual European countries with national currencies. The practical problem of aggregating the components of a macro variable-e.g., EU-wide GDP-turned out to be utterly complex and even more so when data were nonstationary. Juselius and Beyer (2009) studied the sensitivity of different aggregation methods and proposed a procedure that properly accounted for the nonstationarity of the series.

### 3.2. Linking Theory and Evidence: A Bridging Principle

The question of how to link a macroeconomic model to the data is a difficult one. A statistically well-specified empirical model (necessary for correct inference) and an economically well-specified theoretical model represent two basically different entities. To make things worse, econometricians and economists often use concepts which sound similar
but have different meanings. The concepts of exogeneity, steady-state, and equilibrium are just a few examples. Johansen and Juselius (2006) was an attempt to improve the dialog between the econometrician and the economist by offering a dictionary between the two languages.

Based on my experience with CVAR modeling, I became convinced that macroeconomic data were primarily informative about long-run economic relations identified among the cointegrated relations $\beta^{\prime} x_{t}$, and about the exogenous forces, measured by the stochastic trends $\alpha_{\perp}^{\prime} \sum_{i=1}^{t} \varepsilon_{i}$. Recursive constancy tests convinced me that the transitory effects, measured by $\Gamma_{i}$, were inherently unstable. The idea was therefore to assess the economic model in two steps: first by testing its long-run equilibrium structure and, if not rejected, then its short-run adjustment structure conditional on the long-run. Econometrically, such a two-step procedure made sense as the long-run parameter estimates are super-consistent contrary to the short-run which are ordinary consistent.

In 1999 I was invited to give a presentation at a conference on "Macroeconomics and the Real World" held in Bergamo, Italy. At that time I had been struggling to formulate a complete set of testable long-run hypotheses for a model of monetary inflation (Friedman 1970; Romer 1996), subsequently labeled a theory-consistent CVAR scenario. Kevin Hoover, my official discussant, got interested in the idea and we have been collaborating since then. My Bergamo paper was published in the special issue of the Journal of Economic Methodology (Juselius 1999b).

Over the next many years I continued to develop principles for how to translate basic assumptions about the shock structure and steady-state behavior of the monetary model into testable hypotheses on the pulling and pushing forces of the CVAR. Such a theory-consistent CVAR scenario is a summary of the empirical regularities one should find in the data if the basic assumptions of the theoretical model are empirically valid. This idea became a guiding principle of my book (Juselius 2006), in which I demonstrated that essentially all basic assumptions on monetary inflation in Romer (1996) were strongly rejected by the data.

I also tried to formulate a complete set of testable hypotheses about the purchasing power parity (PPP) and the uncovered interest rate parity (UIP). To my surprise, the results were neither straightforward, nor trivial. But, due to other demanding commitments, it took me roughly 10 years until I finally worked out a full theory-consistent CVAR scenario in a chapter of the Handbook of Econometrics (Juselius 2009b). The paper showed that a stationary PPP was empirically inconsistent with observed integration properties of the data, a result that supported the theory of imperfect knowledge economics (Frydman and Goldberg 2007, 2011).

Massimo Franchi visited our department in 2006-2007, and we decided to take a closer look at Ireland (2004) with the title "A method for taking the model to the data". It is a methodological paper in which a real business cycle theory is formulated as a Dynamic Stochastic General Equilibrium model and estimated based on US data. Both the code and the data were available online. Massimo replicated all results of the paper and showed that many key results were empirically fragile. Based on a theory-consistent CVAR scenario, we tested all basic assumptions. They were all rejected and the main conclusions were reversed (Juselius and Franchi 2007).

In 2008, I was guest editor of a special issue for the E-journal Economics with the title Using Econometrics for Assessing Economic Models. See Juselius (2009a). All submitted papers documented lack of support for at least some of the assumptions of the underlying economic model.

In this period, I supervised numerous students and their empirical results were almost without exception similarly disappointing. It was against this background that I wrote Juselius $(2010,2011)$.

### 3.3. Haavelmo's Probability Approach and the CVAR

As mentioned in the introductory section, my most important methodological inspiration came from the Nobel Prize-winning monograph Haavelmo (1944). In particular, Trygve Haavelmo's discussion of statistical inference in economic models based on experimental design data and on non-experimental data was useful for my understanding. In the first case, data are artificially isolated from other influences so that the validity of the ceteris paribus clause is satisfied. In the second case, data are obtained by "passive" observations for which there is no control of the theory that has generated them. Trygve Haavelmo's simple message was that the statistical inference is valid provided the experimental design is valid. The question was then under which conditions this is the case for macroeconomic models. While a prior economic model may or may not be basically correct, it seldom describes data by passive observations very precisely and the ceteris paribus clause is definitely not satisfied. One could asked whether it was at all possible to confront macroeconomic models with our complex economic reality without compromising high scientific standards. Trygve Haavelmo's answer was to introduce the concept of a "design of experiment" for data obtained by passive observations and discuss the validity of inference in that framework. How to construct such a designed experiment was a question that accompanied me in the many years to come.

To ensure valid inference, I thought the statistical model had to be sufficiently general (broad) to represent a set of possible economic models, among which the most relevant one could be selected. In a typical macro situation, there is a variety of models to choose from, but just one data set obtained by passive observations. Therefore, the habit to just assume that the data have been correctly sampled for a preselected model cannot be considered good science: If the statistical model is restricted from the outset in a theoretically prespecified manner, it would be impossible to know which results are true empirical facts and which are due to the assumptions made. ${ }^{3}$ With time, I became ever more convinced that valid inference requires that data are allowed to speak freely about the underlying economic mechanisms and that a key part of the modeling process entails conditioning on important ceteris paribus variables: data by passive observations are never artificially isolated from other factors.

Thus, it seemed mandatory that a probability-based approach to economics should adequately describe all dominant features of economic data in the broad context of a multivariate dynamic macroeconomic model. Juselius (1994) was an early and incomplete attempt to discuss the CVAR model as such a "design of experiment" for data by passive observations. Roughly 20 years later, in connection with the celebration of Haavelmo's centenary birthday, Hoover and Juselius (2015) provided more elaborate arguments for this claim and Juselius (2015) translated one of Haavelmo's own economic models into a theory-consistent CVAR scenario. This is the closest I have come to demonstrating the potential of the CVAR as a design of experiment for data by passive observations.

## 4. Early Applications

In the early years of my academic career, the extant macroeconomic doctrine was strongly influenced by Milton Friedman's monetary theory, which essentially said that money should be controlled in order to control inflation. Friedman's slogan was that "inflation is always and everywhere a monetary problem". What was needed was a monetary authority that was dedicated to keep money supply aligned with the equilibrium level of a money demand relation. My goal was to estimate such a relation for Denmark.

Most attempts to estimate a money demand relation were based on simple regression models, or in some exceptional cases single equation error correction models. I was convinced that the CVAR model would produce much improved estimates and was therefore excited to apply it to Danish data. Some of the results in Johansen and Juselius (1990) also fulfilled my expectations. I found a completely stable money demand relation with

[^2]a plausible coefficient to the cost-of-holding money, measured by the long-short interest rate spread.

Econometrically, the results were straightforward: the trace test suggested that the rank was one, so there was no need to impose (difficult) identifying restrictions on the long-run structure. Economically, some results were plausible: the estimated cointegration relation was directly interpretable as an equilibrium error from a long-run money-demand relation. However, other results were more puzzling. The adjustment coefficients suggested that only money stock was adjusting to deviations in money demand. Hence, monetary shocks had no permanent effect on the system and the exogenous shocks came from aggregate income, the interest on M2, and the long-term bond rate. That cumulated shocks to the interest rates acted as exogenous drivers to the system was against the expectations hypothesis that predicted a stationary interest rate spread.

It was a successful econometric example, but some of the results were economically puzzling. From day one, I learned the hard lesson that the CVAR approach forces you to understand the economic problem in the full context of its system dynamics. Over the next several years I was driven by the urge to better understand why some of the results were so puzzling, making me investigate alternative inflationary transmission mechanisms. This is what the subsequent subsections are about.

### 4.1. Is Inflation a Monetary Phenomenon?

One problem with the Johansen and Juselius (1990) results was that inflation rate was not part of the VAR system. At that time we were not yet aware of the implication of nominal-to-real transformation that the inflation rate should also be included as a system variable. ${ }^{4}$ Perhaps, the puzzling results were due to the missing inflation rate?

As expected, the CVAR extended with the inflation rate produced one additional cointegration relation, identified as a stationary relation among inflation and the two interest rates. The estimated coefficients of the money demand relation were the same as before, which was not surprising as the cointegration property is invariant to extensions of the information set. However, the rest of the results were also very similar: (i) money stock was still purely adjusting, (ii) monetary shocks had no exogenous impact on the system, and (iii) deviations from long-run money demand did not significantly affect the inflation rate. See Juselius (1998a).

The conclusion was that adding inflation to the system did not resolve the empirical puzzle. In terms of the pulling and pushing forces, the results showed almost the opposite of what I had expected: money stock, the short-term interest rate, and inflation rate were purely adjusting and the long-term bond rate and the real GDP represented the exogenous forces. The hypothesis that an empirically stable money-demand relation is a prerequisite for inflation control was, therefore, completely refuted. Juselius (2006) showed that this conclusion-as well as the other results-was robust to extending the sample with 40 quarterly observations.

I began to ponder whether the Danish inflation rate might have been more affected by the actions of the Bundesbank than of the Danish National Bank. As Denmark is a small open economy and Germany is a strong and dominant neighbor, the idea did not seem too far-fetched. Juselius (1996) investigates this hypothesis by analyzing the monetary transmission mechanisms in Germany. Parameter constancy tests revealed a fundamental break in the structure around 1983 and the sample had to be split in two. The results were quite interesting. In the first period, the results seemed to support my prior: a plausible monetary policy rule was identified and inflation was significantly adjusting to it. In the second period, the same policy rule was found but inflation was no longer adjusting to it. I tentatively concluded that financial deregulation and increased globalization were behind the changes in monetary transmission mechanisms.

[^3]This was the first time I obtained results showing that macroeconomic transmission mechanisms might have changed around the mid-eighties. To learn more, I began to study monetary transmission mechanisms more systematically. Juselius (1998b) compared the Danish and German results with similar analyses of Spain and Italy. While the conclusion was that monetary transmission mechanisms had changed, the results showed that the changes took place at different time points due to different institutional set-ups. The comparative study was, therefore, followed up with more detailed country-specific analyses: Juselius (1998a) discussed the Danish case, Juselius (2001) the Italian case, and Juselius and Toro (2005) the Spanish case.

My many attempts to estimate monetary transmission mechanisms made me increasingly skeptical about Friedman's strong claim. Rather than (CPI) inflation always and everywhere being a monetary problem, the results indicated almost the opposite that inflation was "never and nowhere a monetary problem". ${ }^{5}$

### 4.2. Is Inflation Imported?

The next question, whether Danish inflation is primarily imported, led me to study the international transmission mechanisms between Denmark and Germany. The analysis was motivated by the two theoretical cornerstones of international macroeconomics: the purchasing power parity (PPP) and the uncovered interest rate parity (UIP). The PPP condition was assumed to hold as a stationary or near $I(1)$ process, whereas the UIP condition described a market clearing condition. I found essentially no empirical support for the stationarity of the two conditions: the deviations from the PPP and the UIP exhibited a pronounced persistence that was empirically indistinguishable from a first-or even a second-order nonstationary process, whereas a combination of the two was found to be stationary.

During my work on the PPP-UIP problem, it dawned on me that the CVAR model with its informationally rich pulling and pushing structures contained an enormous potential for combining deductive and inductive inference. Juselius (1995) reports not just tests of the stationarity of the PPP, UIP, and combined relation, but of basically every possible hypothesis related to the foreign transmission mechanisms. This detailed analysis offered a wealth of new information, again some of it quite puzzling. For example, the trace test found the data vector to be $I(2)$ and tests of unit vectors in $\beta$ found prices and the exchange rate to be individually $I(2)$. The test of overall long-run proportionality of the two prices was accepted, whereas proportionality between relative prices and the nominal exchange rate was strongly rejected.

To shed light on this puzzle, I checked the estimates of the stochastic $I(2)$ trend and its loadings. The former showed that the $I(2)$ trend was primarily generated by the twice cumulated shocks to the long-term German bond rate. The latter showed that the $I(2)$ trend loaded onto the two prices but also onto the exchange rate, explaining the lack of cointegration between the price differential and the nominal exchange rate. The fact that the stochastic $I(2)$ trend originated from shocks to the German bond rate and that it loaded onto the nominal exchange rate (as well as onto the Danish and German price levels) pointed to the financial market as a crucial player in the foreign exchange market. Roman Frydman and Michael Goldberg pointed out to me that the results were consistent with imperfect knowledge expectations in a monetary model for exchange rate determination. It was the beginning of a long collaboration between Roman and Michael and the econometrics group in Copenhagen.

In 1996, Søren and I moved to the European University Institute in Florence, Italy, for five years. Ronald McDonald was a visiting scholar during this period and we initiated a joint collaboration on the PPP and UIP for USA-Germany and USA-Japan. The information set was now extended with the short-term interest rates and we used monthly rather

[^4]than quarterly observations. The inclusion of short rates in the analysis allowed us to additionally address the expectations hypothesis and the term structure of interest rates. While this increased the richness of the economic structures, extending the system to seven equation seriously complicated the identification of the long-run cointegration structure. The solution was first to analyze a smaller model-consisting of prices, the long-term interest rates, and the nominal exchange rate-and then to use the cointegration results of the smaller model as the starting point for the big model. This procedure-dubbed specific-to-general in the choice of the information set-builds on the invariance of cointegration to expansions of the information set. ${ }^{6}$ If cointegration is found in a smaller set of variables, it will also be found in an extended set. Since then, I have successfully used this principle as a means to manage long-run identification in high-dimensional systems. (Juselius 2006, chp. 19) provides a detailed discussion of the merits of this method.

The research results were published in Juselius and MacDonald $(2004,2006)$. Many of the findings were similar to the ones in Juselius (1995). Long-run proportionality between the price differential and the nominal exchange rate was also now strongly rejected for both country pairs. But, unlike Juselius (1995), we applied the nominal-to-real transformation, nonetheless, and performed the analysis in the $I(1)$ model, acknowledging the loss of some data information. ${ }^{7}$

The results showed that inflation rates were, again, purely adjusting, so inflationary shocks had no long-run effect on the system. An interesting result was the very slow inflation adjustment to the PPP, in contrast to the fast adjustment to the combined PPP-UIP relation. It suggested that the long and persistent deviations from PPP were sustainable as long as they were compensated by similar deviations in the interest rate differential. The long-term bond rates were found to be weakly as well as strongly exogenous in both the small and the big system. Interestingly, the real exchange rate was weakly exogenous in the small system but no longer so in the big system. Thus, statistically significant adjustment of the real exchange rate required the short rates to be part of the model, illustrating the peril of the ceteris paribus clause for conclusions when data are non-stationary.

At that time, many of the results were puzzling based on standard theory: (i) inflationary shocks were not driving nominal interest rates, instead interest rate shocks were pushing the inflation rates; (ii) the long-term bond rates were exogenous to the system rather than the short rates; and (iii) the short-long interest spread was nonstationary in contrast to the expectations hypothesis. While these results were puzzling from the point of view of standard macroeconomic models, Juselius (2017) subsequently showed that they were perfectly consistent with the theory of Imperfect Knowledge Economics (Frydman and Goldberg 2007, 2011).

### 4.3. CPI Inflation and Excessive Wage Claims

While Juselius (1995) showed that Danish inflation was partly imported, the extent to which wage inflation had been pushing price inflation was still an open question. My first study of wage, price, and unemployment dynamics is described in (Juselius 2006, chp. 20). The choice of variables, manufacturing wages, consumer prices, producer prices, productivity, and unemployment was motivated by standard theories for centralized wage bargaining, assuming that a proposed pay rise by the labor union reflects a trade-off between a higher consumption wage against lower employment. Whether the employers' union accepts the pay rise is assumed to be a trade-off between future profits and firm

[^5]competitiveness against the increased risk of a union strike. Both unions are assumed to maximize their share of future productivity increases.

During the sample period (1971:1-2003:1) the European markets had become increasingly integrated implying on one hand improved profit possibilities, on the other more fierce competition. For Danish enterprises, facing relatively high wage costs, the latter was a serious problem. The consequence of the almost fixed krone in the EMS arrangement after 1983 was that a less competitive export firm could no longer count on exchange rate realignments to improve its competitiveness. To remain in the market, an exporting enterprise had basically three possibilities: (i) to reduce employment until the marginal cost equaled the competitive price, (ii) to increase labor productivity, or (iii) to outsource production. All three measures were used and all of them affected the unemployment rate.

From the eighties onward, unemployment rates fluctuated in long and persistent swings around long-run average values, not just in Denmark but in most European countries. These long and persistent unemployment episodes were puzzling from the point of view of standard theories that assumed unemployment rates to be stationary around a constant rate, the natural rate of unemployment. This inspired Edmund Phelps to write the theory of "Structural Slumps" (Phelps 1994), arguing that the natural rate of unemployment is a function of the real interest rate and/or the real exchange rate.

These considerations motivated me to extend the data vector with the long-term bond rate and the real exchange rate. The system, now containing seven variables, was quite large and I used the specific-to-general approach in the choice of information set to manage the complexity of identifying a plausible long-run structure. In the first step, I analyzed the first five of the set of variables and, in the second step I added the interest rate, and in the third step the real exchange rate. This allowed me to study the effect of the ceteris paribus assumption "real interest rate and real exchange rate constant" on wage determination. It also allowed me to test some of the fundamental hypotheses of Phelps' structural slumps theory and helped me to understand how globalization and financial deregulation had affected the mechanisms of the labor market.

The results showed that the nominal wage and the two price variables were individually $I(2)$ and that overall long-run homogeneity among them was statistically acceptable. Therefore, based on the nominal-to-real transformation, the nominal variables were replaced by the real consumer wage, the price wedge between consumer and producer wages, and consumer price inflation. Based on this change, the model could now be analyzed in the $I(1)$ framework without loss of information. The econometrically motivated price wedge was also an important economic variable, as its coefficient can be interpreted as a measure of the relative bargaining power of employers and employees. The price wedge is also assumed to reflect the degree of product market competition, which-if high-is likely to result in pricing-to-market behavior (Krugman 1986).

The empirical results of the Danish wage and price mechanisms are discussed in detail in (Juselius 2006, chp. 20). One important finding-revealed by the tests of parameter constancy-was a significant change in the mechanisms around mid-eighties. The change was so fundamental-similar to the German monetary mechanisms in 1983-that it left me with no other options than to split the sample period in two parts: the first part comprised the seventies up to mid-eighties, the other from the mid-eighties up to 2003.

The results for the first regime suggested a narrative that was about strong labor unions, rigid institutions, devaluations and realignments and, for the second regime, about increasingly weak labor unions and improvements of labor productivity. Excessive wage claims seemed to have caused both price inflation and unemployment in the first regime but foremost unemployment in the second. In the second regime, competitiveness was largely achieved by producing the same output with less labor as evidenced by unemployment and trend-adjusted productivity being cointegrated. There was evidence of a Phillips curve relationship in both regimes, but it was rather insignificant in the first whereas strongly significant in the second. In the latter regime, the strong co-movements between
unemployment and the real bond rate were consistent with a Phelpsian natural rate. In both regimes, inflation was significantly adjusting to the real exchange rate.

I found the results exciting and was eager to know whether they had any generality outside Denmark. At this time, Javier Ordonez visited our department and we decided to study the Spanish wage and price dynamics using a similar approach. The Spanish results, published in Juselius and Ordóñez (2009), showed that the basic mechanisms behind the determination of wage, price and unemployment were very similar, albeit with some differences that seemed to reflect institutional differences between the two countries. In a recent article, Juselius (2021b) find support for the above mechanisms based on US data.

### 4.4. Combining the Results: A Proposal for a Large-Scale Macro Model

The advantage of the VAR approach is that the data are allowed to speak freely without being silenced by prior restriction. The disadvantage is that the number of parameters increases substantially with each included variable. Adding one variable leads to $(2 p-1) k$ new parameters, where $p$ is the dimension of the variable vector and $k$ is the autoregressive lag. This can quickly become prohibitive in macroeconomic models, where sample periods seldom are very long.

To circumvent this problem, Juselius (1992) proposed a procedure for combining partial models into a larger macro model. The idea was to study how CPI inflation was affected by monetary inflation, wage inflation, and imported inflation by estimating cointegration relations in three partial VAR models. Econometrically, the procedure is based on the invariance of the cointegration property to expansions of the information set. Economically, it rests on the interpretation of a properly identified cointegration relation as a deviation from a long-run equilibrium value, implying that it could be treated as a convenient summary measure of the most important information from the sector in question. For example, if wages at time $t$ are on the equilibrium level, then the value of the cointegration relation would be approximately zero, implying no wage pressure on CPI inflation. In contrast, if the absolute value of the cointegration relation is large, then wages are either below or above their equilibrium level with a potentially large impact on CPI inflation.

I used the same idea in Juselius (2006, Part VI) where more detailed and extensive analyses of the three sectors are reported. Figure 1 below illustrates the procedure. First, the relevant long-run relations are identified based on smaller CVAR models, then the deviations from these relations enter as the main explanatory variables in a bigger model explaining key economic determinants, such as CPI inflation, the unemployment rate and the interest rate ${ }^{8}$. The list of key variables can of course be extended as illustrated in Juselius (2006, chp. 22). For the period 1972-2003, the results showed that (i) the identified cointegration relations represented the major bulk of the explanatory power with only minor effects from short-run changes of the system variables, (ii) excess money had essentially no effect on the CPI inflation rate, (iii) wage inflation had a large inflationary effect until capital deregulation in the mid-eighties and only a modest effect afterwards, and (iv) wage increases reflected a smaller part of the productivity growth after globalization and capital deregulation than before. Labor unions seemed to have become increasingly powerless.

While not perfect, the results from the big combined model seemed very promising. The idea of using the "specific-to-general" in the choice of information set and the "general-to-specific" in the search for a parsimoniously parametrized model begun to look like a feasible way to overcome both the dimensionality problem of the CVAR and the complexity problem of large macro models.

[^6]

Figure 1. Using cointegration relations from partial models as the main economic determinants in a large macroeconomic model.

I was excited about the possibility to use the above principle to handle large scale macro models, such as the traditional Keynesian macro models consisting of numerous behavioral relations in which endogeneity, exogeneity, and ceteris paribus are given a priori. Such a behavioral relation could be subject to a CVAR analysis without the need to fix the status of a variable as endogenous or exogenous and with the possibility to add relevant ceteris paribus variables. Furthermore, the stationarity of the presumed behavioral relations could be properly tested and efficiently estimated, dynamic feedback effects and long-run dynamic multiplier effects would be readily available.

By combining such partial dynamic models into a large-scale model of the economy, one would obtain something resembling a general (dis)equilibrium macromodel. It would be based on the assumption that deviations from equilibrium values-the equilibrium errors-are the most crucial determinants of key variables in the economy, such as output growth, unemployment, wage inflation, interest rate, CPI inflation, house price inflation, stock price inflation, and real exchange rate. At the same time it would provide useful information about the dynamics of each subsector of the economy. I thought it would give large-scale macro models a much needed face lift and be a powerful method for an improved understanding of our complex economic reality. To my disappointment the idea has not yet been realized anywhere in the world, at least not to my knowledge.

## 5. Towards a New Methodological Approach

After having applied the CVAR to numerous empirical problems, it became ever more evident that there was more persistence in the data than standard models could explain. I often found the data to be indistinguishable from $I(2)$ and this was not just for price variables, like the CPI, but also for relative prices, nominal and real exchange rates, and even real and nominal interest rates, which a priori were expected to be stationary or at most $I(1)$. Even unemployment, another important real economy variable, was often found to be indistinguishable from $I(2)$ and cointegrated with the real interest rate and the real exchange rate.

Many economists would argue that such findings are implausible as economic variables could not drift away forever as a true $I(2)$ process can, nor could equilibrium errors be $I(1)$ as economic variables do not move infinitely away from their equilibrium values. However, while this is obviously correct, it does not exclude the possibility that variables over finite samples may exhibit a persistence that is empirically indistinguishable from
a unit root or a double unit root process. Furthermore, because economic relationships seldom remain unchanged for long periods of time, the infinity argument may not be very relevant in economics. In line with this, Juselius (2013) argues that a statistical unit root should not be given an interpretation as a structural economic parameter and that the classification of variables/relations as either stationary, (near) $I(1)$ or (near) $I(2)$ is a requisite for successful empirical modeling.

What makes a near $I(2)$ process extremely interesting is that such a process is able to generate long-lasting swings, a typical feature of economic variables (Johansen 1997 2006; Paruolo and Rahbek 1999). In spite of this, applications of the $I(2)$ model are rare in the literature. To understand why, Juselius (2014) discusses a simple case, $\Delta x_{t}=\omega_{t}+\varepsilon_{x, t}$ where $\omega_{t}=\omega_{t-1}+\varepsilon_{\omega, t}$ and the shocks $\varepsilon_{\omega, t}$ are small compared to the shocks $\varepsilon_{x, t}$ i.e., the signal-to-noise ratio is small. Simulations show that univariate Dickey-Fuller tests hardly ever detect the second unit root in the drift term, whereas the multivariate tests almost always find it. This is particularly so when the signal-to-noise-ratio is small, typical of asset prices in speculative markets. As most people use univariate rather than multivariate tests to determine the order of integration, the results may explain why econometricians/economists find economic variables/relations to be $I(1)$ rather than $I(2)$.

Why is this important? Knowing the approximate order of integration and cointegration among variables is a very important and useful piece of information in the modelling process. For example, an $I(1)$ variable cannot be significantly related to an $I(0)$ variable, neither can an $I(2)$ variable to an $I(1)$ variable, but they can be combined to form a stationary cointegrated relationship. Therefore, by exploiting the information in the data given by the integration/cointegration properties of the variables, one can obtain robust estimates of long-run, medium-run, and short-run structures in the data, thus improving the specification of the economic model. In the words of Hoover et al. (2009), the CVAR allows the data to speak freely about the mechanisms that have generated them. Juselius $(2006,2013)$ provide more detailed discussions.

### 5.1. Long Swings in Financial Market Behavior

At that time, financial behavior was rarely part of macroeconomic models as-somewhat simplistically -a fully rational financial actor was assumed to know when the market price deviated from its equilibrium price and then would act accordingly. Rational financial markets would, therefore, drive financial prices back to equilibrium and the equilibrium prices would correctly reflect movements in the real economy. Because financial prices were assumed to be correctly determined, deregulated financial markets were good, not harmful, to the real economy. Therefore, there was no need to regulate and no reason to worry about the effect of financial market behavior in macroeconomic models. The reasoning relied on the efficient market hypothesis, that was based on the rational expectations hypothesis and the assumption that economic models are known and stable over time. However, all these assumptions seemed at odds with what I constantly saw in the data: the frequent structural breaks, the frequent changes of exogeneity status, the long and persistent swings around equilibrium values indistinguishable from a unit root process.

That the deviations from some of the fundamental economic parities-the Fisher parity, the term spread, the purchasing power parity, the uncovered interest rate paritywere statistically indistinguishable from unit root processes seemed particularly worrisome to me. Where did this additional persistence come from? It seemed inconsistent with standard REH models which assumed much faster adjustment to long-run equilibria. Why did the persistent swings not vanish with the nominal-to-real transformation when the nominal deflator was the consumer price index? It gradually dawned on me that long and persistent swings in both the nominal and the real magnitude of a variable were typically found in prices associated with financial behavior, such as exchange rates, interest rates, stock prices, house prices, energy prices, and prices for precious metals. It raised the question why did they fluctuate in a manner detached from the development of standard consumer prices and real productivity growth in the economy?

As already mentioned, this empirically very strong feature turned out to be largely consistent with a monetary model for the exchange rate based on imperfect knowledge expectations (Frydman and Goldberg 2007, 2011). The imperfect knowledge argument is that no one can know-not even in probabilistic terms-what the true fundamental value of an asset is. This is because the value of a financial asset is a function of future-unpredictable-cash flows. Given such Knightian uncertainty, market participants interpret in diverse ways a wide range of news about fundamental factors, from real growth and inflation rate announcements to political developments and debt crises. This diversity combined with loss aversion can then explain why forecasts of future asset prices tend to generate persistent movements around benchmark values.

The theory of imperfect knowledge economics provided me with an explanation of the puzzling finding that the real exchange rate and the real interest rate differential were empirically near $I(2)$. In Frydman et al. $(2008,2012)$ we addressed the PPP puzzle and the long swings puzzle both theoretically and empirically.

Another strain was offered by Hommes (2006) and Hommes et al. (2005a, 2005b), which similarly focus on the persistent swings in asset prices. In this theory, financial markets are populated by fundamentalists using economic fundamentals to forecast future price movements, and by chartists-trend-followers-using technical trading rules to forecast prices. Financial actors are switching endogenously between mean-reverting fundamentalists and trend-following chartists depending on how far away the price is from long-run equilibrium values. Positive feedback prevails when the chartists dominate the market and negative feed-back when the fundamentalists dominate.

Common to the above models is that today's asset price depends on future prices which, in varying degree, are being forecasted under imperfect knowledge and, therefore, deviate from the price derived under the REH. In both models prices can deviate from longrun benchmark values for extended periods of time, thereby generating self-reinforcing expectational cycles. All this seemed to provide a rational for my puzzling findings and was a motivation to focus on financial behavior and its role for the real economy.

### 5.2. Persistent Movements and Time-Varying Coefficients

How to analyze such self-reinforcing expectational cycles econometrically is, however, far from simple. Inspired by (Frydman and Goldberg 2007, 2011), Juselius and Assenmacher (2017) interpreted the long swings in the real US dollar-Swiss franc rate in the context of a simple model with time-varying coefficients using the following assumptions: A financial actor understands that PPP holds in the long run, but not necessarily in the short run. $\mathrm{He} /$ she is, therefore, in the short-term likely to react on a number of other determinants, $z_{t}$, such as changes in interest rates, relative incomes and consumption, and many more. In such a world, financial actors tend to attach time-varying weights, $B_{t}$, to relative prices depending on how far away the nominal exchange rate is from its fundamental PPP value, i.e.,

$$
\begin{equation*}
s_{t}=A+B_{t}\left(p_{d, t}-p_{f, t}\right)+z_{t} \tag{3}
\end{equation*}
$$

where $s_{t}$ is the log of the nominal exchange rate, $p_{d, t}-p_{f, t}$ is the $\log$ of the relative price between domestic and foreign country, and $B_{t}$ fluctuates around 1.0. The change in the nominal exchange rate can then be expressed as

$$
\Delta s_{t}=B_{t} \Delta\left(p_{d, t}-p_{f, t}\right)+\Delta B_{t}\left(p_{d, t}-p_{f, t}\right)-\Delta B_{t} \Delta\left(p_{d, t}-p_{f, t}\right)+\Delta z_{t}
$$

where $\Delta B_{t} \Delta\left(p_{d, t}-p_{f, t}\right)$ can be assumed very small. In addition, Frydman and Goldberg (2007) makes the assumption that $\left|\Delta B_{t}\left(p_{d, t}-p_{f, t}\right)\right| \ll\left|B_{t} \Delta\left(p_{d, t}-p_{f, t}\right)\right|$. This is backed up by simulations showing that a change in $\Delta B_{t}$ has to be implausibly large for $\Delta B_{t}\left(p_{d, t}-p_{f, t}\right)$ to have a noticeable effect on $\Delta s_{t}$. Therefore,

$$
\begin{equation*}
\Delta s_{t} \simeq B_{t} \Delta\left(p_{d, t}-p_{f, t}\right)+\Delta z_{t} \tag{4}
\end{equation*}
$$

where $\Delta s_{t}, \Delta\left(p_{d, t}-p_{f, t}\right)$ and $\Delta z_{t}$ are typically near $I(1)$ processes. To study the properties of this type of time-varying parameter model, Tabor (2014) considered the CVAR model:

$$
\begin{align*}
\Delta Y_{t} & =\alpha\left(Y_{t-1}-\beta_{t} X_{t-1}\right)+\varepsilon_{y, t}, \text { where } \varepsilon_{y, t} \sim \operatorname{niid}\left(0, \sigma_{\varepsilon_{y}}^{2}\right)  \tag{5}\\
\Delta X_{t} & =\varepsilon_{x, t} \text { where } \varepsilon_{x, t} \sim \operatorname{niid}\left(0, \sigma_{\varepsilon_{x}}^{2}\right)
\end{align*}
$$

He generates the data with $\alpha=-1$ and $\beta_{t}=\beta_{0}+\rho \beta_{t-1}+\varepsilon_{\beta, t}$, where $\varepsilon_{\beta, t} \sim$ $\operatorname{niid}\left(0, \sigma_{\varepsilon_{\beta}}^{2}\right)$. Then, $E\left(\beta_{t}\right)=\frac{\beta_{0}}{1-\rho}=\beta$ for $\rho=\{0.0,0.5,0,95\} . \alpha=-1$ implies that the adjustment of $Y_{t}$ back to $\beta_{t}^{\prime} X_{t}$ is immediate. Instead of estimating a time-varying parameter model, Morten fitted a constant parameter CVAR model to the simulated data, so that $\left(\beta_{t}-\beta\right) X_{t}$ becomes part of the CVAR residual. The results in Tabor (2014) show that the closer $\rho$ is to 1 , the more persistent is the estimated gap term, $Y_{t}-\hat{\beta}^{\prime} X_{t}$, and the smaller is the estimated adjustment coefficient $\alpha$-albeit still highly significant. Furthermore, as long as $\rho<1$, the mean of the estimated $\hat{\beta}$ approximately equals its true value $\beta$. When $\rho=1$, this is no longer the case.

Thus, the pronounced persistence away from long-run equilibrium values and the small adjustment coefficients often found in constant-parameter CVAR models is potentially a result of time-varying coefficients due to forecasting under imperfect knowledge. Juselius (2017) shows that this may explain the persistence of the PPP gap and the inability to reject $I(2)$ persistence using the CVAR. Even though under this assumption, the $I(2)$ model is just an approximation of a model with time-varying coefficients, it may, nonetheless, be a useful approximation. The linear VAR with constant parameters gives access to a vast econometric literature on estimation and testing, whereas the complexity of estimating a time-varying parameter VAR model is daunting except for small models with only one or a few time-varying parameters.

When analyzing the PPP and the UIP conditions for various countries based on $I(2)$ CVAR models, the results frequently supported the main assumption of the imperfect knowledge based monetary model that the deviations from the PPP was cointegrated with the spread between the domestic and foreign real interest rates. By interpreting the persistent movements in the real exchange rate as a proxy for an uncertainty premium in the foreign currency market-proposed by Frydman and Goldberg (2007)—the results show strong empirical support for a stationary uncertainty adjusted UIP condition. Furthermore, Johansen et al. (2010) reported an econometric analysis of the full set of international parity conditions using German-US data.

Juselius and Assenmacher (2017) also report a similar study of Swiss-US data in which equilibrium error-increasing behavior is used to identify the channels through which self-reinforcing feedback mechanisms takes place. The results show that such behavior plays a significant role for the persistent fluctuations in exchange rates, interest rates, and prices. They also show that once loss-aversion and uncertainty is allowed for, the excess return puzzle disappears, suggesting that agents are behaving rationally but that imperfect knowledge outcomes are very different from the ones in an REH world.

### 5.3. Real Exchange Rate Persistence and the Real Economy

The derived CVAR scenario for an imperfect knowledge monetary model in Juselius (2017) provides an explanation for why asset prices, but not CPI prices, tend to fluctuate in long persistent swings and, consequently, why real trend-adjusted asset prices are empirically almost indistinguishable from their nominal trend-adjusted magnitudes, and why interest rate differentials are near $I(2)$. The imperfect knowledge-based monetary model fits the data remarkably well as shown in Juselius (2006, chp. 21) for Denmark versus Germany, Juselius and MacDonald (2004) for Japan versus USA, Juselius and MacDonald (2006) for Germany versus USA, Juselius and Assenmacher (2017) for Switzerland versus USA, and Juselius and Juselius and Stillwagon (2018) for UK versus USA.

Common to the above papers is the finding that Purchasing Power Parity needs Uncovered Interest Parity to become a stationary parity relation. The implication is that
equilibrium in the goods market is not directly associated with purchasing power parity but with a stationary relation between a nonstationary real exchange rate and the interest rate spread. Thus, the real exchange rate can persistently appreciate/depreciate as long as the real interest rate differential moves in an offsetting manner. As these persistent swings around equilibrium values are caused by speculative behavior in the market for foreign exchange, they are essentially outside domestic policy control-at least as long as transactions in the foreign currency market are neither regulated nor taxed.

For the US dollar and the UK pound market, Juselius and Stillwagon (2018) found that it is the interest rate expectations-measured by consensus forecasts of professional forecasters-that are pushing the interest rates and the exchange rate in the long run. Furthermore, the results show that it is the shocks to the US consensus forecasts-rather than the UK ones-that are dominating the long persistent swings. An interesting finding is that changes in the nominal exchange rate are pushing the foreign currency market in the medium run with interest rates following suit, whereas expectational shocks to the interest rates are pushing the market in the long run with the nominal exchange following suit. These results are basically consistent with imperfect knowledge based models.

That the fundamental parity conditions-in particular the PPP and the UIP-were systematically found to be non-stationary, prompted the question of how this is affecting the real economy.

Juselius (2013) was my first attempt to address the two-way interdependence between the real economy and financial behavior in asset markets. The theme of the paper was strongly influenced by Phelps' hypothesis that the natural rate of unemployment is a function of the real interest rate and/or the real exchange rate. Because Phelps' "Structural Slumps" book assumed the latter to be stationary, I was excited to examine the implications of them being nonstationary instead.

In a stationary world, exporting and importing enterprises would be insulated from changes in the relative costs, if the nominal exchange rate correctly reflects relative costs between the two countries. In a nonstationary world, where the nominal exchange rate is typically determined by speculative transactions, it is much less affected by the trade in exports and imports. ${ }^{9}$ Thus, an exporting firm would have to resort to "pricing-to-market" strategies rather than mark-up-pricing (Krugman 1986), or it would lose market shares. For example, over a prolonged period of currency appreciation, such a firm will experience a mounting pressure to be competitive. As raising the price is not feasible, there are few other options than to improve productivity. This can be done, for example, by requiring workers to produce more per hour, firing the least productive workers, outsourcing production, or introducing new technology. All these measures affect unemployment rate. When the exchange rate finally reverses-now depreciating-the pressure on competitiveness is released but, because competing enterprises in foreign countries now experience an appreciating exchange rate and, therefore, have to resort to similar measures, prices do not rise much.

Thus, consumer prices-determined by fierce competition in an international marketremain low and stable, whereas asset prices-determined by speculative expectationstend to fluctuate in long persistent swings. The fact that the unemployment rate and trendadjusted productivity have been co-moving and that the natural rate of unemployment has been a function of the real interest rate-rather than a constant-are consistent with the above mechanisms (Juselius 2006, chp. 20; Juselius and Ordóñez 2009).

These results can also explain the inflation puzzle, e.g., why inflation has been low and stable over time (below $2 \%$ for several decades) at the same time as the nominal interest rate has moved in long persistent swings and, hence, why CPI inflation and nominal interest rate are typically not found to be cointegrated, against the Fisher parity assumption.

[^7]They also suggest that any attempt to control inflation by changing central bank interest rate is likely to be ineffective. To be effective, such a policy rule would require the above parities to hold as stationary conditions, or important parts of the transmission mechanism are broken. Evidence of this was found in Johansen and Juselius (2001) in which the Federal Funds rate was shown to be an inefficient instrument for US inflation control during the Greenspan monetary policy period. While the inflation rate has been low in periods of inflation targeting, my claim is-supported among others by the CVAR analyses in Juselius (1998b)—that it has been so for other reasons, primarily financial deregulation and global competition.

One consequence of the low inflation rate is that the pressure on the central bank to raise its interest rate has been low for several decades. Exceptionally low interest rate levels have in turn led to easy credit and a corresponding strong increase in liquidity. The consequence of high, credit-financed, demand for real estate and stock is that house and stock prices have sky-rocketed. At the same time, the CPI inflation rate has remained low. Juselius (2019) reports a comprehensive analysis of the soaring Danish house and stock prices, totally detached from the CPI prices and the real GDP, that ultimately led to the Danish house price bubble in 2007 and then to the financial crisis in 2008.

Juselius (2019) demonstrates empirically that accruing imbalances often tend to counterbalance each other, sometimes over extended periods of time and argues that a balance maintained by several imbalances is a very fragile balance: sooner or later a large shock to the system will cause the balance to collapse-as happened in 2007 when the house price bubble burst and similarly in 2008 when the financial crisis hit the world economy with unprecedented force. Thus, the great recession seems to have grown out of many imbalances-initiated by financial behavior-which were allowed to develop over a long time.

Over time I have become ever more convinced that financial behavior is an extremely important determinant of the real economy. This was also the main conclusion in Colander et al. (2009) which already in 2008 argued that unrealistic financial models have had a large and detrimental effect on real economies. A few months later, this claim turned out to be almost too correct.

### 5.4. Crises Periods and Comparative Studies

At a time when many argued that the Great Recession was a once in a life time event-a black swan-that could not have been foreseen, I vividly remembered a similar crisis at the beginning of the nineties in Finland. The deregulation of the Finnish credit market in 1986 had resulted in an overheated economy and in strongly increasing real estate prices. When the house price bubble burst, unemployment rates soared and reached more than $20 \%$-from a starting position of $1.6 \%$-in a very short period of time. In a joint project with my son Mikael Juselius (Juselius and Juselius 2013), we asked the questions (i) whether the Finnish experience could be understood as a balance sheet recession ${ }^{10}$, (ii) whether the unemployment dynamics made sense in the context of Phelps' Structural Slumps theory (Phelps 1994), and (iii) whether the theory of Imperfect Knowledge Economics (Frydman and Goldberg 2007, 2011) could explain the persistent movements in the data. To answer these questions, we applied the CVAR model to inflation, unemployment, a short- and a long-term interest rate.

Econometrically, our CVAR model performed surprisingly well—considering the wild fluctuations of the Finnish data. The results-reported in Juselius and Juselius (2013)—gave support to all three priors: the Phelps' hypothesis that the natural rate of unemployment is a function of the real interest rate; the Frydman and Goldberg Imperfect Knowledge hypothesis of pronounced persistence in the long-term real interest rate; and the Koo hypothesis of the Central Bank interest rate as an ineffective instrument during a balance

[^8]sheet recession. Furthermore, based on a smooth transition model in which the transition variable was designed to capture household sector leverage-adjusted for movements in the value of the housing collateral, the paper demonstrated that strongly increasing house prices had played a crucial role for the depth and the length of the subsequent crisis. As soon as house prices started falling and the house debt exceeded the value of the collateral, the leverage effect was shown to become extremely important.

Altogether, the Finnish results seemed to be able to shed light on the dynamic transmission mechanisms of inflation, unemployment and interest rates in a crisis period. It raised the question whether the results could have been used to foresee the housing bubble 15 years later, or whether there were lessons to be learned for other countries with a similar bubble experience. The latter was the underlying motivation in Juselius and Dimelis (2019) to address the empirical mechanisms governing the Greek depression, the most serious and destructive of all European crises.

Many aspects of the Finnish crisis were similar to the ones in Greece: the deregulation of the Finnish credit market in 1986 resulted in a booming housing market and a serious house price bubble; joining the eurozone caused the Greek bond rate to drop to previously unprecedented levels and caused a credit financed boom in aggregate demand. As in Finland, Greek wages and prices-in particular real estate prices-were rising and competitiveness was deteriorating. When the Greek bubble burst, the drop in aggregate income and the rise in unemployment rate were huge and of similar magnitudes as in Finland. However, the Greek crisis, while similar in many aspects to the Finnish one, differs strongly in others. For example, the source of the debt (private/public, external/internal), the strong/weak institutional set-up, and in particular the exchange rate regime are defining differences of crucial importance. The fact that Finland was able to devalue its currency while Greece was not, is likely to have made all the difference for the length of the crisis. It is one reason why the comparison with Finland is interesting.

Unlike the Greek economy, Finland managed to get out of the crisis in approximately three-admittedly very hard-years by devaluing the Finnish markka with $33 \%$. Moreover, unlike the Greek experience, the Finnish unemployment rate came down quite fast, albeit stabilizing at a somewhat higher level compared to the pre-crisis period. One reason why the Greek unemployment was stuck at very high levels seemed to be the prolonged period of policy uncertainty following the outbreak of the crisis. Unlike the Finnish analysis, the Greek analysis therefore required a variable measuring confidence as well as two variables measuring the development of the Greek competitiveness within and outside the eurozone.

In the Greek analysis, the most striking result was a critical relationship between the bond rate and the unemployment rate: As the crisis erupted, the bond rate increased sharply followed by a strong increase in unemployment, the increase in unemployment rate caused the bond rate to increase further and unemployment to follow suit, and so on. This vicious cycle was orchestrated by a continuous fall in the confidence rate that kept deteriorating until relative producer costs stopped increasing around 2012. The empirical results showed that all variables, except CPI inflation, exhibited self-reinforcing feedback behavior somewhere in the system, a feature that is likely to have aggravated the problems and effectively prevented good policy solutions. As the euro rate was determined by factors mainly outside the Greek control, Greece was stuck in a situation with no feasible options: a dramatic lowering of wage costs was politically impossible; leaving the euro would have been extremely costly due to the large external debt. At the same time the confidence in the Greek economy continued to drop which added to the depressed state of the economy.

The two papers illustrate an important methodological principle: by using the same "experimental design", here the CVAR model, and controlling for institutional differences by conditioning on appropriately selected variables, one can learn about similarities and dissimilarities in different economies. I thought this would be particularly valuable when addressing policy changes and the response to them.

Juselius et al. (2014) followed a similar principle when studying the effectiveness of foreign aid in 36 South Saharan African countries. Among these, 29 countries were classified into four more homogeneous groups according to aid effectiveness ${ }^{11}$. Within these groups more detailed analyses were performed. The results showed that the overall qualitative conclusions were rather similar for the vast majority of South Saharan African countries. However, the results also showed that the dynamics of the transmission of aid onto the macro economy differed a lot depending on the classification. This was not surprising as such: aid is given for different purposes in different countries. Econometrically, our results pointed to the peril of using panel data analyses as a basis for policy advice in such diverse countries as South Saharan African countries. As aid effectiveness has frequently been studied based on panel data analyses which-implicitly or explicitly-assume homogeneous countries across the panel, this should be a reason for concern.

A small number of countries fell outside the classification criteria, among them Ghana and Tanzania. Based on an extended data set, Juselius et al. (2017) studied the transmission mechanisms of aid in more detail for these two countries. It turned out that both countriesfor political reasons-had manipulated their exchange rate for extended periods of time, with the consequence that the aid transmission mechanism did not follow a standard pattern. Conditional on the anomalous exchange rate regimes, the aid transmission results became economically interpretable again. Yet another example of the importance of the ceteris paribus clause.

The three papers in this section and many others mentioned earlier serve the purpose of illustrating the potential of the CVAR as a design of experiment for data obtained by passive observations. Perhaps it is time to challenge the frequent claim that it is not possible to apply designed experiments in macroeconomics.

## 6. Some Reflections

The title of this paper "Searching for a theory that fits the data" was chosen to emphasize the distinction between my own empirical approach and the one that underpins most empirical research in economics: "Searching for data that fits the theory". This difference reflects, no doubt, what the researcher considers most important: the empirical reality or the theory supposed to explain it. For me it was never a choice: to better understand what was going on in the empirical economy was the main reason why I chose a university career in economics. To develop empirical methods that could increase the transparency of economic mechanisms and potentially improve economic policy decisions has been an important personal driver in all these years of extremely hard work.

To stumble over the CVAR and see its great potential as a methodology for empirical economics was like winning a lottery. However, while I believe the CVAR has fully lived up to its promises, the way it has been applied in the literature has been disappointing. Numerous papers report all kind of CVAR analyses, most of them give the impression of being done by statistical non-experts: data have been read in and the CVAR button has been pushed. However, the CVAR methodology cannot be applied mechanically: it depends upon the researcher's statistical expertise and requires interaction between the econometrician, the economist and the data. For example, it does not make sense to work with a CVAR model until you have checked whether (1) the sample period is representative for your questions, (2) the chosen information set is sufficiently broad to answer the questions of interest, (3) the most important institutional changes have been controlled for, (4) the parameters of interest are reasonably stable over time, (5) the residual mis-specification tests are acceptable, and many more. If you sidestep these important steps, you will very likely get nonsense. Perhaps this is the reason why the impact of the CVAR on economic modeling has been so disappointing.

[^9]A frequent claim is that the quality and informational content of macroeconomic data are too low for a CVAR analysis to be reliable. I agree that economic time series data seldom represent the true measurements of the theoretical model. For example, the representative agent's income, consumption, and hours worked in a DSGE model has little in common with the various measurements of aggregate income, private consumption, and total hours worked that can be found in the national statistical publications. However, while macro data are clearly contaminated with measurement errors, such errors may not be of great concern for the more important long-run analysis unless they are systematic and cumulate to a nonstationary process. Whatever the case, theoretically correct measurements do not exist and, thus, cannot be used by politicians and decision makers to react on. The forecasts, plans and expectations that agents base their decisions on are the observed data and we better understand them, however imperfect they are.

A related claim is that, unless the empirical model is constrained by theory from the outset, one would not be able to make sense of the results: without the mathematical logic of the theoretical model, one opens up for quackery. I hold the opposite view. Scientific objectivity can only be achieved provided data are not constrained from the outset in a theoretically prespecified direction. When they are, it is impossible to know which results are due to the assumptions made and which are true empirical findings. This point was amply illustrated in Juselius and Franchi (2007) where essentially all assumptions underlying a DSGE model by Ireland (2004) were found to lack empirical support. When a well-specified CVAR was fitted to the same data the results showed that all conclusions were reversed. Thus, the conclusions of the Ireland paper reflect the assumptions made rather than true empirical findings.

Another related claim is that CVAR models are so general that they can show anything. ${ }^{12}$ Over time, I have applied the CVAR model to numerous problems in a variety of countries and for many different time periods. These applications have convinced me that macroeconomic data are surprisingly informative, but only if you let them speak freely about the story they want to tell. This, of course, does not mean that data can speak by themselves without theory, nor without rigor: a CVAR analysis should obey equally strict rules as a mathematical analysis of an economic model. A well-specified CVAR model, estimated by a full information maximum likelihood method, describes by definition all aspects of the data and, thus, summarizes the empirical features that an empirically relevant theory should be able to explain. Typical features are unit root nonstationarity, structural change, non-constant parameters, dynamic long-run equilibrium relationships, and selfreinforcing feedback mechanisms. All of them have strong implications for the choice of economic model. For example, $I(2)$ nonstationarity is consistent with economic relations that deviate persistently-in a near $I(1)$ manner-from long-run equilibria, suggesting that the choice of economic model should be based on disequilibrium economics. Guzman and Stiglitz (2020) discuss the basic features of such a theory.

Disequilibrium economics again points to complex adjustment dynamics and nonstandard—non-REH—expectations as the relevant concepts. Data covering crises periods typically reveal such features. While many economists consider crisis periods to be aberrations outside the range of economic modeling, they are not outside the range of a well-specified CVAR analysis. Economic crises are often devastating for ordinary people's lives and any lesson that can be learned should not be missed. In 2008, standard mainstream models did not spot the accruing imbalances and, hence, failed to prevent and explain the economic crisis. As these models are still based on essentially the same assumptions, it seems unlikely that they will be able to foresee the next crisis in time to prevent it. See also the critique in Stiglitz (2018).

[^10]The CVAR results have typically favored traditional Keynesian macroeconomics, albeit modified with expectations based on uncertainty and imperfect knowledge and controlling for the effect of financial behavior. As the dynamic macroeconomic disequilibrium theory proposed in Guzman and Stiglitz (2020) is broadly in line with the above, I argue in Juselius (2021b) that the CVAR may have the potential to work as an empirical methodology for disequilibrium macroeconomics. The fact that the many-then theoretically puzzling-CVAR results reported in Section 4 would no longer be puzzling in the context of disequilibrium economics, should contribute to the plausibility of this suggestion.

I will end the tale of my personal odyssey by hoping that this journey, bending and looping as it has been, can convince at least some econometricians, economists and policymakers that well-founded empirical findings rather than theoretical convictions should guide economic policy. The abundance of theoretically puzzling-but empirically and econometrically well founded results-signal the need for new theory and deserve to be taken seriously. No doubt, empirically unfounded economic policy is likely to have exacerbated some of the defining problems of our time, such as recurring crises, increasing inequality, and growing populism. The development of a more relevant macroeconomics that serve not just the few but all is desperately needed.

Funding: This research received no external funding.
Institutional Review Board Statement: Not applicable.
Informed Consent Statement: Not applicable.
Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Acknowledgments: I would like to thank the guest editors, Rocco Mosconi and Paolo Paruolo, for asking me to prepare a survey paper of my past and ongoing research. I was very happy to accept it. In all these years, the Danish Social Sciences Research Council has generously supported my research with several grants. In the last years prior to my retirement, I was fortunate to obtain a Centre grant from the Institute of New Economic Thinking. I am thankful for all the support I obtained. I am also grateful for valuable comments from two anonymous referees, from the two guest editors, and from Michael Goldberg and Roman Frydman.

Conflicts of Interest: The author declares no conflict of interest.

## References

Colander, David, Michael Goldberg, Armin Haas, Katarina Juselius, Alan Kirman, Thomas Lux, and Birgitte Sloth. 2009. The Financial Crisis and the Systemic Failure of the Academics Profession. Critical Review 21: 249-67. [CrossRef]
Dennis, Jonathan, Henrik Hansen, Søren Johansen, and Katarina Juselius. 2006. CATS in RATS. Cointegration Analysis of Time Series. Version 2. Evanston: Estima.
Doornik, Jonathan, and Katarina Juselius. 2017. Cointegration Analysis of Time Series Using CATS 3 for OxMetrics. New York: Timberlake Consultants Ltd.
Engle, Robert, David Hendry, and Jean-Francois Richard. 1983. Exogeneity. Econometrica 51: 277-304. [CrossRef]
Friedman, Milton. 1970. The Counterrevolution in Monetary Theory. Occasional paper 33. London: Institute of Economic Affairs.
Frydman, Roman, and Michael Goldberg. 2007. Imperfect Knowledge Economics: Exchange Rates and Risk. Princeton: Princeton University Press.
Frydman, Roman, and Michael Goldberg. 2011. Beyond Mechanical Markets: Risk and the Role of Asset Price Swings. Princeton: Princeton University Press.
Frydman, Roman, Michael D. Goldberg, Søren Johansen, and Katarina Juselius. 2008. A Resolution of the Purchasing Power Parity Puzzle: Imperfect Knowledge and Long Swings. Working Paper 31. Copenhagen: Economics Department, University of Copenhagen.
Frydman, Roman, Michael Goldberg, Søren Johansen, and Katarina Juselius. 2012. Long Swings in Currency Markets: Imperfect Knowledge and I(2) Trends. Copenhagen: University of Copenhagen.
Granger, Clive. 1983. Cointegrated Variables and Error Correction Models. Discussion Paper. San Diego: University of California.
Guzman, Martin, and Joseph Stiglitz. 2020. Towards a Dynamic Disequilibrium Theory with Randomness. Working Paper 27453, Cambridge: National Bureau of Economic Research.
Haavelmo, Trygve. 1944. The Probability Approach to Econometrics. Econometrica 12: 1-118. [CrossRef]
Hansen, Henrik, Søren Johansen, and Katarina Juselius. 1994. CATS in RATS. Version 1.0. In Manual to Cointegration Analysis of Time Series. Evanstone: ESTIMA.

Hendry, David, and Katarina Juselius. 2000. Explaining Cointegration Analysis. Part 1. The Energy Journal 21: 1-42. [CrossRef]
Hendry, David, and Katarina Juselius. 2001. Explaining Cointegration Analysis. Part 2. The Energy Journal 22: 1-52. [CrossRef]
Hommes, Cars. 2006. Heterogeneous Agent Models in Economics and Finance. In Handbook of Computational Economics 2. Edited by Leigh Tesfation and Kenneth Judd. Amsterdam: Elsevier.
Hommes, Cars, Hai Huang, and Duo Wang. 2005a. A Robust Rational Route to Randomness in a Simple Asset Pricing Model. Journal of Economic Dynamics and Control 29: 1043-72. [CrossRef]
Hommes, Cars, Joep Sonnemans, Jan Tuinstra, and Henk van der Velden. 2005b. Coordination of Expectations in Asset Pricing Experiments. Review of Financial Studies 18: 955-80. [CrossRef]
Hoover, Kevin, Søren Johansen, and Katarina Juselius. 2009. Allowing the Data to Speak Freely: The Macroeconometrics of the Cointegrated Vector Autoregression. American Economic Review 98: 251-55. [CrossRef]
Hoover, Kevin D., and Katarina Juselius. 2015. Trygve Haavelmo's Experimental Methodology and Scenario Analysis in a Cointegrated Vector Autoregression. Econometric Theory 31: 249-74. [CrossRef]
Ireland, Peter. 2004. A Method for Taking Models to the Data. Journal of Economic Dynamics \& Control 28: 1205-26.
Johansen, Søren. 1988. Statistical Analysis of Cointegration Vectors. Journal of Economic Dynamics and Control 12: 231-54. [CrossRef]
Johansen, Søren. 1992. Cointegration in Partial Systems and the Efficiency of Single Equation Analysis. Journal of Econometrics 52: 389-402. [CrossRef]
Johansen, Søren. 1994. The Role of the Constant and Linear Terms in Cointegration Analysis of Non-Stationary Variables. Econometric Reviews 13: 205-30. [CrossRef]
Johansen, Søren. 1996. Likelihood-Based Inference in Cointegrated Vector Autoregressive Models. Oxford: Oxford University Press.
Johansen, Søren. 1997. Likelihood Analysis of the I(2) Model. Scandinavian Journal of Statistics 24: 433-62. [CrossRef]
Johansen, Søren. 2006. Statistical Analysis of Hypotheses on the Cointegrating Relations in the I(2) Model. Journal of Econometrics 132: 81-115. [CrossRef]
Johansen, Søren, and Katarina Juselius. 1990. Maximum Likelihood Estimation and Inference on Cointegration-With Applications to the Demand for Money. Oxford Bulletin of Economics and Statistics 52: 169-210. [CrossRef]
Johansen, Søren, and Katarina Juselius. 1992. Testing Structural Hypotheses in a Multivariate Cointegration Analysis of the PPP and the UIP for UK. Journal of Econometrics 53: 211-244. [CrossRef]
Johansen, Søren, and Katarina Juselius. 1994. Identification of the Long-Run and Short-Run Structure: An Application to the ISLM Model. Journal of Econometrics 63: 7-36. [CrossRef]
Johansen, Søren, and Katarina Juselius. 2001. Controlling Inflation in a Cointegrated Vector Autoregressive Model with an Application to US Data. Working Paper 3, Copenhagen: University of Copenhagen.
Johansen, Søren, and Katarina Juselius. 2006. Extracting Information from the Data: A European View on Empirical Macro. In Post Walrasian Macroeconomics: Beyond the Dynamic Stochastic General Equilibrium Model. Edited by David Colander. Cambridge: Cambridge University Press, pp. 301-33.
Johansen, Søren, and Katarina Juselius. 2014. An Asymptotic Invariance Property of the Common Trends under Linear Transformations of the Data. Journal of Econometrics 178: 310-15. [CrossRef]
Johansen, Søren, Katarina Juselius, Roman Frydman, and Michael D. Goldberg. 2010. Testing Hypotheses in an I(2) Model with Piecewise Linear Trends. An Analysis of the Persistent Long Swings in the Dmk/\$Rate. Journal of Econometrics 158: 117-29. [CrossRef]
Johansen, Søren, and Anders Rygh Swensen. 1999. Testing Exact Rational Expectations in Cointegrated Vector Autoregressive Models. Journal of Econometrics 93: 73-91. [CrossRef]
Johansen, Søren, and Anders Rygh Swensen. 2004. More on Testing Exact Rational Expectations in Cointegrated Vector Autoregressive Models: Restricted Constant and Linear Term. The Econometrics Journal 7: 389-97. [CrossRef]
Juselius, Katarina. 1992. Domestic and Foreign Effects on Prices in an Open Economy: The Case of Denmark. Journal of Policy Modeling 14: 401-28. [CrossRef]
Juselius, Katarina. 1993. VAR Modelling and Haavelmo's Probability Approach to Macroeconomic Modelling. Empirical Economics 18: 595-622. [CrossRef]
Juselius, Katarina. 1994. On the Duality Between Long-Run Relations and Common Trends in the $I(1)$ versus $I(2)$ Model: An Application to Aggregate Money Holdings. Econometric Reviews 13: 151-79. [CrossRef]
Juselius, Katarina. 1995. Do Purchasing Power Parity and Uncovered Interest Rate Parity Hold in the Long Run? An Example of Likelihood Inference in a Multivariate Time-Series Model. Journal of Econometrics 69: 211-40. [CrossRef]
Juselius, Katarina. 1996. An Empirical Analysis of the Changing role of German Bundesbank after 1983. Oxford Bulletin of Economics and Statistics 58: 791-817. [CrossRef]
Juselius, Katarina. 1998a. A Structured VAR for Denmark under Changing Monetary Regimes. Journal of Business and Economic Statistics 16: 400-11.
Juselius, Katarina. 1998b. Changing Monetary Transmission Mechanisms within the EU. Empirical Economics 23: 455-81. [CrossRef]
Juselius, Katarina. 1999a. Price Convergence in the Long Run and the Medium Run. An $I(2)$ analysis of six price indices. In Cointegration, Causality, and Forecasting. Festschrift in Honour of Clive Granger. Edited by Robert Engle and Hal White. Oxford: Oxford University Press.
Juselius, Katarina. 1999b. Models and Relations in Economics and Econometrics. Journal of Economic Methodology 6: 259-90. [CrossRef]

Juselius, Katarina. 2001. European Integration and Monetary Transmission Mechanisms: The Case of Italy. Journal of Applied Econometrics 16: 341-58. [CrossRef]
Juselius, Katarina. 2006. The Cointegrated VAR Model: Methodology and Applications. Oxford: Oxford University Press.
Juselius, Katarina. 2009a. Special Issue on Using Econometrics for Assessing Economic Models: An Introduction. Economics: The Open-Access, Open-Assessment E-Journal 3: 1-28. [CrossRef]
Juselius, Katarina. 2009b. The Long Swings Puzzle: What the Data Tell When Allowed to Speak Freely. In Palgrave Handbook of Econometrics: Vol. 2: Applied Econometrics. Edited by Kerry Patterson and Terence Mills. London: Palgrave Macmillan, pp. 349-84.
Juselius, Katarina. 2010. Time to Reject the Privileging of Economic Theory over Empirical Evidence? A Reply to Lawson. Cambridge Journal of Economics 35: 423-36. [CrossRef]
Juselius, Katarina. 2011. On the Role of Theory and Evidence in Macroeconomics. In The Elgar Companion to Recent Economic Methodology. Edited by Wade Hands and John Davis. Cheltenham: Edward Elgar Publishing, pp. 404-26, chap. 17.
Juselius, Katarina. 2013. Imperfect Knowledge, Asset Price Swings and Structural Slumps: A Cointegrated VAR Analysis of Their Interdependence. In Rethinking Expectations: The Way Forward for Macroeconomics. Edited by Edmund Phelps and Roman Frydman. Princeton: Princeton University Press, pp. 328-50.
Juselius, Katarina. 2014. Testing for Near I(2) Trends When the Signal to Noise Ratio is Small. Economics: The Open-Access, OpenAssessment E-Journal 8: 1-30. [CrossRef]
Juselius, Katarina. 2015. Haavelmo's Probability Approach and the Cointegrated VAR. Econometric Theory 31: 213-32. [CrossRef]
Juselius, Katarina. 2017. Using a Theory-Consistent CVAR Scenario to Test an Exchange Rate Model Based on Imperfect Knowledge. Econometrics 5: 30. [CrossRef]
Juselius, Katarina. 2019. Økonomien og Virkeligheden: et Opgør med Finanskapitalismen. Copenhagen: Informations Forlag.
Juselius, Katarina. 2021a. A Theory-Consistent CVAR Scenario: Testing a Rational Expectations Based Monetary Model. Econometrics, forthcoming
Juselius, Katarina. 2021b. Disequilibrium Macroeconometrics. In Industrial and Corporate Change on Macro and Development. Edited by Joseph Stiglitz and Giovanni Dosi. Forthcoming
Juselius, Katarina, and Katrin Assenmacher. 2017. Real Exchange Rate Persistence and the Excess Return Puzzle: The Case of Switzerland versus the USA. Journal of Applied Econometrics 32: 1145-55. [CrossRef]
Juselius, Katarina, and Andreas Beyer. 2009. Does it Matter how to Measure Aggregates? Monetary Transmission Mechanisms in the Euro Area. In The Methodology and Practice of Econometrics: A Festschrift in Honour of David Hendry. Edited by Jennifer Castle and Neil Shephard. Oxford: Oxford University Press, pp. 365-85.
Juselius, Katarina, and Sophia Dimelis. 2019. The Greek Crisis: A Story of Self-Reinforcing Feed-Back Mechanisms. Economics: The Open-Access, Open-Assessment E-Journal 13: 1-22. [CrossRef]
Juselius, Katarina, and Massimo Franchi. 2007. Taking a DSGE Model to the Data Meaningfully, Economics, The Open-Access, OpenAssessment E-Journal 1. [CrossRef]
Juselius, Katarina, and Mikael Juselius. 2013. Balance Sheet Recessions and Time-Varying Coefficients in a Phillips Curve Relationship: An Application to Finnish Data. In Essays in Nonlinear Time Series Econometrics. Edited by Niels Haldrup, Mika Meitz and Pentti Saikkonen. Oxford: Oxford University Press.
Juselius, Katarina, and Ronald MacDonald. 2004. International Parity Relationships Between the USA and Japan. Japan and the World Economy 16: 17-34. [CrossRef]
Juselius, Katarina, and Ronald MacDonald. 2006. International Parity Relationships and a Nonstationary Real Exchange Rate: Germany versus the USA in the Post Bretton Woods Period. In International Macroeconomics: Recent Developments. Edited by Morales Zumaquero. New York: Nova Publishers, pp. 79-103.
Juselius, Katarina, Niels Framroze Møller, and Finn Tarp. 2014. The Long-Run Impact of Foreign Aid in 36 African Countries: Insights from Multivariate Time Series Analysis. Oxford Bulletin of Economics and Statistics 76: 153-84. [CrossRef]
Juselius, Katarina, and Javier Ordóñez. 2009. Balassa-Samuelson and Wage, Price and Unemployment Dynamics in the Spanish Transition to EMU Membership. Economics: The Open-Access, Open-Assessment E-Journal 3. [CrossRef]
Juselius, Katarina, Abdulaziz Reshid, and Finn Tarp. 2017. The Real Exchange Rate, Foreign Aid and Macroeconomic Transmission Mechanisms in Tanzania and Ghana. Journal of Development Studies 53: 1075-103. [CrossRef]
Juselius, Katarina, and Juan Toro. 2005. Monetary Transmission Mechanisms in Spain: The Effect of Monetization, Financial Deregulation, and the EMS. Journal of International Money and Finance 24: 509-31. [CrossRef]
Juselius, Katarina, and Joshua Stillwagon. 2018. Are Outcomes Driving Expectations or the Other Way Around? An I(2) CVAR Analysis of Interest Rate Expectations in the Dollar/Pound Market. Journal of International Money and Finance 83: 93-105. [CrossRef]
Kongsted, Hans Christian. 2005. Testing the Nominal-to-Real Transformation. Journal of Econometrics 124: 205-25. [CrossRef]
Koo, Richard. 2010. The Holy Grail of Macroeconomics: Lessons from Japan's Great Recession. Hoboken: John Wiley \& Sons.
Krugman, Paul. 1986. Pricing to Market when the Exchange Rate Changes. Working Paper 1926, Cambridge: National Bureau of Economic Research.
Paruolo, Paolo, and Anders Rahbek. 1999. Weak Exogeneity in I(2) VAR Systems. Journal of Econometrics 93: 281-308. [CrossRef]
Phelps, Edmund. 1994. Structural Slumps. Princeton: Princeton University Press.
Romer, David. 1996. Advanced Macroeconomics. New York: McGraw Hill.

Sargan, Dennis. 1964. Wages and Prices in the United Kingdom: A Study of Econometric Methodology. In Econometric Analysis for National Economic Planning. Edited by Peter Hart, Riley Mills and John Whitaker. London: Butterworth Co., pp. 25-63.
Stiglitz, Joseph. 2018. Where Modern Macroeconomics Went Wrong. Oxford Review of Economic Policy 34: 70-106.
Tabor, Morten N. 2014. Time-Varying Cointegration Parameters as a Source of Slow Error-Correction in the Cointegrated VAR Model. Essays on Imperfect Knowledge Economics, Structural Change, and Persistence in the Cointegrated VAR Model. Ph.D thesis, University of Copenhagen, Copenhagen, Denmark.
econometrics

## Article

# Søren Johansen and Katarina Juselius: A Bibliometric Analysis of Citations through Multivariate Bass Models 

Fragiskos Archontakis ${ }^{1, *}$ and Rocco Mosconi ${ }^{2}$<br>1 School of Economics, Business Administration \& Legal Studies, International Hellenic University, 57001 Thermi, Greece<br>2 Dipartimento di Ingegneria Gestionale Politecnico di Milano, Piazza L. da Vinci 32, 20133 Milano, Italy; rocco.mosconi@polimi.it<br>* Correspondence: f.archontakis@ihu.edu.gr

Citation: Archontakis, Fragiskos, and Rocco Mosconi. 2021. Søren Johansen and Katarina Juselius: A Bibliometric Analysis of Citations through Multivariate Bass Models. Econometrics 9: 30. https://doi.org/ 10.3390/econometrics9030030

Academic Editor: In Choi

Received: 22 April 2021
Accepted: 10 August 2021
Published: 12 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.


Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/).


#### Abstract

We showcase the impact of Katarina Juselius and Søren Johansen's contribution to econometrics using bibliometric data on citations from 1989 to 2017, extracted from the Web of Science (WoS) database. Our purpose is to analyze the impact of KJ and SJ's ideas on applied and methodological research in econometrics. To this aim, starting from WoS data, we derived two composite indices whose purpose is to disentangle the authors' impact on applied research from their impact on methodological research. As of 2017, the number of applied citing papers per quarter had not yet reached the peak; conversely, the peak in the methodological literature seem to have been reached around 2000, although the shape of the trajectory is very flat after the peak. We analyzed the data using a multivariate dynamic version of the well known Bass model. Our estimates suggest that the methodological literature is mainly driven by "innovators", whereas "imitators" are relatively more important in the applied literature: this might explain the different location of the peaks. We also find that, in the literature referring to KJ and SJ , the "cross-fertilization" between methodological and applied research is statistically significant and bi-directional.


Keywords: bass diffusion model; bibliometrics; cointegration

## 1. Introduction

Using bibliometric methods in order to value the quantity and quality of knowledge produced by researchers is increasingly the standard practice in most disciplines (Garfield et al. 1978; Redner 1998). In the field of economics, Kalaitzidakis et al. (1999) provided a ranking of European departments based on ten top journals, which was later updated and expanded to include, amongst others, also a ranking of academic journals in economics (Kalaitzidakis et al. 2003). At the same time, Coupé (2003) published a paper including rankings for researchers based on publications and citations; there, he explicitly mentions the highly-cited work by Søren Johansen and Katarina Juselius on cointegration, stating that "first in the citation ranking is Søren Johansen. Thanks to his top cited papers on cointegration written at the beginning of the 1990s, he is first on the three different citation rankings"(Coupé 2003, p. 1336).

The aim of this paper is to showcase, through a bibliometric analysis, the impact of Katarina Juselius (KJ) and Søren Johansen's (SJ) contribution to the field of econometrics. An important distinctive trait of their scientific production is to combine methodological and applied research, placing their work in the so-called "Pasteur's Quadrant" (Stokes 1997), characterized by use-inspired basic research, where applied objectives are chased in parallel with fundamental scientific creativity. This motivates our main research question: what is the influence of KJ and SJ's work on applied and methodological research in econometrics? We believe that, from this analysis, we can learn something about the mechanisms of scientific discovery in general. Although the methodology used in this paper is different, our analysis has some resemblance to the work by Stigler (1994), who analyzed citation data
in the journals of statistics and probability, investigating the mechanisms of knowledge diffusion within and across fields. Among other findings, he observed that "there is a tendency for influence to flow from theory to applications to a much greater extent than in the reverse direction" (see Stigler 1994, p. 94). As we will show in this paper, this is, to some extent, confirmed also in the abundant literature inspired by KJ and SJ, although the flow running from applied econometrics toward econometric methodology is also clearly visible in this case. We think that this depends on the peculiar approach to empirical research inspired by KJ and SJ work, "in which data would be allowed to speak freely without being silenced by prior restriction and in which basic hypotheses could be adequately tested and empirically relevant structures estimated"-the quote is taken from Juselius (2021, p. 6), in the same Special Issue of Econometrics hosting this paper. This approach requires a continuous dialogue with methodologists, posing to them challenging requests for appropriate statistical models and suitable probability results allowing for correct inference within such models.

Our empirical investigation is based on citation data collected through the Web of Science (WoS) database, based on which we derived two new composite indices whose purpose is to disentangle the citations originated in the applied econometric research from those coming from the methodological research. Our analysis reveals that the majority of citations (about $85 \%$ ) arise from applied research. Of course, to put this figure into perspective, one should compare it with the share of methodological research in econometrics in general: unfortunately, we do not have this information (our impression is that the share is somewhat lower than $15 \%$ ). Interestingly, the dynamic pattern of the two indices is quite different: the citation peak in the applied literature does not seem to be reached yet, whereas the peak in the methodological literature seems to have occurred around the turn of the century. To analyze these bibliometric data, we resorted to a multivariate dynamic version of the well known Bass (1969) model, proposed by Boswijk et al. (2009) building on Franses (2003), Boswijk and Franses (2005) and Fok and Franses (2007). Bibliometric evidence suggests that Bass-type models provide a useful way to fit most Nobel in Economics prize winner citation trajectories; see Bjork et al. (2014). This fact might indicate that, up to a point, economic knowledge could follow the well-known product life cycle, which is usually characterized by the following phases: introduction, growth, maturity (including peak) and decline, within the context of a scholar's professional lifetime and beyond. An interesting aspect of Bass models is that they describe the diffusion pattern as dependent on two key parameters, $p$ and $q$, measuring the relevance of innovation and imitation, respectively: these two parameters are shown in Min et al. (2018) to have an important role in the growth and decay of citation counts in several scientific disciplines. In this paper, we will show that, according to our estimates, the relative importance of imitative and innovative mechanisms is quite different for methodological and applied econometric research: this seems to be responsible for the different trajectories of the two research strands.

The paper is organized as follows: Section 2 describes the data collection and management process to support the analysis. Section 3 presents the univariate and multivariate Bass model, and Section 4 illustrates our empirical findings. Finally, Section 5 concludes and provides directions for further research.

A word on notation used in the paper. The backshift operator $L$ is defined as $L X_{t}=X_{t-1}$, where $X_{t}$ is a time series; the difference operator is defined as $\Delta=1-L$, so that $\Delta X_{t}=X_{t}-X_{t-1}$. $\boldsymbol{I}_{n}$ is the $n \times n$ identity matrix, $\boldsymbol{u}_{n, i}$ is the $i$-th column of $\boldsymbol{I}_{n}$, $\mathbf{1}_{n}=\sum_{i=1}^{n} \boldsymbol{u}_{n, i}, \mathbf{0}_{m, n}$ is an $m \times n$ matrix of zeros, $\operatorname{diag}\left\{\boldsymbol{a}_{i}\right\}$ is the block diagonal matrix whose generic diagonal block is the matrix $a_{i}$ (of course any of the $\boldsymbol{a}_{i}$ 's could also be a scalar).

## 2. The Data

This section describes the line of thought and the data collection process, including the source and sample size, while providing some preliminary analysis through stylized facts.

To the purpose of this study, we consider the scientific production on cointegration by KJ and SJ as an indissoluble whole, where economic questions motivate the development of econometric theory and the development of econometric theory sharpens the economic questions. Their papers on cointegration are, therefore, analyzed together, whether single authored or coauthored and whether the main focus is on methodology (with just an illustrative example) or application (with a pedagogical effort to illustrate how the methodology can be applied to a real problem).

On 9 April 2018, we collected from the Web of Science (WoS) the data about the citations received by KJ and SJ for papers between 1989:Q1 and 2017:Q3. ${ }^{1}$ For practical reasons, we limit the analysis to the 10 most quoted papers, which are presented in Table 1, sorted by publication date.

Table 1. The 10 "Top Cited" papers by S. Johansen and K. Juselius, in chronological order (data collected on 9 April 2018).

| Order (Time) | Paper | Citations (WoS) | New Citations (WoS) |
| :---: | :---: | :---: | :---: |
| 1 | Johansen (1988) | 4008 | 4008 |
| 2 | Johansen and Juselius (1990) | 2567 | 1060 |
| 3 | Johansen (1991) | 2256 | 997 |
| 4 | Johansen (1992c) | 170 | 45 |
| 5 | Johansen and Juselius (1992) | 477 | 90 |
| 6 | Johansen (1992b) | 249 | 40 |
| 7 | Johansen (1992a) | 251 | 69 |
| 8 | Johansen and Juselius (1994) | 196 | 32 |
| 9 | Johansen et al. (2000) | 167 | 60 |
| 10 | Hendry and Juselius (2001) | 112 | 56 |
|  | 10,453 | 6457 |  |

The total number of citations received by the top ten papers amounted at that time to 10,453 , whereas the number of citing papers was $6457,{ }^{2}$ so that every citing paper cites, on average, 1.62 papers, with a maximum of 7 observed four times. ${ }^{3}$ In terms of the number of citations, well ahead of the rest of the publications, are the papers by Johansen (1988) in the Journal of Economic Dynamics \& Control with 4008 citations, the joint paper by Johansen and Juselius (1990) in the Oxford Bulletin of Economics \& Statistics with 2567 citations, and the paper by Johansen (1991) in Econometrica with 2256 citations. For each paper, the last column in Table 1, "new citations", indicates the number of citing papers referring to that paper and to none of the earlier ones: for example, paper number 7 is cited by 251 papers, but only 69 of them cite paper number 7 and none of the earlier. A high "new citations"/ citations ratio suggests that the paper has broken new ground in the field: for example, the paper by Hendry and Juselius (2001) treats data from the field of energy, and as a result, energy-related papers often cite Hendry and Juselius (2001), rather than the earlier papers. Notice that the first three papers account for $84.5 \%$ of the citations and $93.9 \%$ of the citing papers.

To avoid double counting, we focus on the number of citing papers rather than on the number of citations, ${ }^{4}$ and we define by $c_{t}$ the number of citing papers published in quarter $t$ ( $t$ ranges from 1989:Q1 to 2017:Q3, i.e., 115 quarters). Based on the WoS data, we split $c_{t}$ into two composite indices aimed at measuring the impact of KJ and SJ ideas on applied and methodological econometric research, respectively. ${ }^{5}$ To this aim, we have analyzed each of the 6457 citing papers, classifying them according to their methodological or applied nature. The classification is essentially based on the title of the citing paper. ${ }^{6}$ We adopted the following classification:

- Purely applied (PA) papers: the title refers to an application, with no reference to an econometric method, technique or issue. We have found $n_{P A}=4198$ such papers.
- Mainly applied (MA) papers: the title refers both to an econometric method, technique or issue and an application, and the focus seems to be on the latter (e.g., "Does exchange-rate volatility affect import flows in G-7 countries? Evidence from cointegration models"). We have found $n_{M A}=1451$ such papers.
- Purely methodological (PM) papers: the title refers to an econometric method, technique or issue, with no reference at all to an application. We have found $n_{P M}=716$ such papers.
- Mainly methodological (MM) papers: the title refers both to an econometric method, technique or issue and an application, and the focus seems on the first (e.g., "Robust cointegration testing in the presence of weak trends, with an application to the human origin of global warming"). We have found $n_{M M}=92$ such papers.
We have, therefore, derived four quarterly time series, labeled $c_{P A, t}, c_{M A, t}, c_{P M, t}$ and $c_{M M, t}$, counting the citing papers of each group in each quarter; of course, $c_{t}=$ $c_{P A, t}+c_{M A, t}+c_{P M, t}+c_{M M, t}$. The four time series are reported in Figure 1, where one can observe that the behavior of $c_{P A, t}$ and $c_{M A, t}$ is quite similar, steadily increasing over time, with some low frequency fluctuations, which seem to be shared by both series (the correlation is $69.5 \%$ ). Conversely, $c_{P M, t}$ has a peak around the year 2000 with about 10 papers per quarter, and then it declines until 2005 , seeming to stabilize at around 5 papers per quarter. The series $c_{M M, t}$ is irregular, due to the small number of MM papers, but resembles $c_{P M, t}$ to some extent, as it shows a higher frequency around the year 2000; then, the frequency seems to slightly decline.


Figure 1. Time series plot of $c_{P A, t}, c_{M A, t}, c_{P M, t}$ and $c_{M M, t}$, quarterly data from 1989:1 to 2017:3.
Finally, by combining the four series with suitable weights, we obtained two composite indicators, whose purpose is to measure the impact of KJ and SJ ideas on applied $\left(c_{1, t}\right)$ and methodological ( $c_{2, t}$ ) research:

$$
\begin{align*}
& c_{1, t}=c_{P A, t}+\omega c_{M A, t}+(1-\omega) c_{M M, t}  \tag{1}\\
& c_{2, t}=c_{P M, t}+\omega c_{M M, t}+(1-\omega) c_{M A, t} \tag{2}
\end{align*}
$$

Of course $c_{1, t}+c_{2, t}=c_{t}$ for any $\omega$ by construction. Composite indicators have several pros and cons, as illustrated for example in Nardo et al. (2008) and Kuc-Czarnecka et al. (2020): they allow to summarize complex, multi-dimensional realities, reducing the dimensionality. On the other hand, they might simplify too much, and, even more importantly, the selection of indicators and weights could be the subject of dispute. It is, therefore, important to motivate clearly one's weighting choice and to provide an extensive sensitivity analysis. We provide a thorough discussion of both aspects in Appendix B. In short, the baseline results presented in this paper are based on $\omega=0.85$. This choice is motivated by two main reasons: (i) $\omega$ should be in the range from 0.5 to 1 , extremes excluded, since the papers classified as MA (or MM) should contribute mainly ( $\omega>1 / 2$ )
to the applied (or methodological) index but also, to a lesser extent ( $1-\omega>0$ ), to the methodological (or applied) index; (ii) $\omega=0.85$ would be approximately equal to $n_{P A} /\left(n_{P A}+n_{P M}\right)=0.8545$-in practice, this corresponds to the assumption that the share of "applied research" of an MA paper is similar, on average, to the share of applied research in the econometric literature referring to KJ and SJ papers in general. Notice, however, that, as illustrated in Appendix B, the main results of our econometric analysis are robust to the choice of $\omega$ in the range from 0.5 to 1 .

In order to fix ideas, we provide a short example based on the first few citations in the WoS data. For the year 1989, we have to split the only two existing citations by Baillie and Bollerslev (classified as MA) and Gilbert (classified as PM). ${ }^{7}$ Thus, for this example we have obtained the series illustrated in Table 2.

Table 2. Illustration of the classification scheme: citations in 1989.

| Author(s) | Quarter | $c_{t}$ | $\boldsymbol{c}_{P A, t}$ | $\boldsymbol{c}_{\boldsymbol{M A}, \boldsymbol{t}}$ | $\boldsymbol{c}_{\boldsymbol{P M}, \boldsymbol{t}}$ | $\boldsymbol{c}_{\boldsymbol{M M}, \boldsymbol{t}}$ | $\boldsymbol{c}_{1, t}$ | $\boldsymbol{c}_{2, t}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Baillie-Bollerslev | 1989:Q1 | 1 | 0 | 1 | 0 | 0 | 0.85 | 0.15 |
|  | 1989:Q2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Gilbert | 1989:Q3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
|  | 1989:Q4 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |

The cumulative applied index at time $T=115$, i.e., 2017:Q3, is equal to the following:

$$
\begin{aligned}
C_{1, T} & =\sum_{t=1}^{T}\left(c_{P A, t}+\omega c_{M A, t}+(1-\omega) c_{M M, t}\right) \\
& =4198+0.85 \times 1451+0.15 \times 92=5445.2
\end{aligned}
$$

while the cumulative methodological index will be equal to the following:

$$
\begin{aligned}
C_{2, T} & =\sum_{t=1}^{T}\left(c_{P M, t}+\omega c_{M M, t}+(1-\omega) c_{M A, t}\right) \\
& =716+0.85 \times 92+0.15 \times 1451=1011.8
\end{aligned}
$$

This shows that the majority of citations originates from applied research: defining $C_{t}=C_{1, t}+C_{2, t}$, the ratio $\frac{C_{1, T}}{C_{T}}$ is $84.4 \%$, whereas $\frac{C_{2, T}}{C_{T}}$ is $15.6 \%$. To check the appropriateness of our classification scheme, we analyzed how these ratios vary by publishing journal. Tracking down the 6457 citing papers, we obtained from the WoS database that they appeared in 696 distinct journals. Table 3 provides the ranked list of the top 20 journals by the number of citing papers: these journals hosted 2676 citing papers, i.e., $41.4 \%$.

The evidence in Table 3 seems to confirm the validity of our classification: the average $c_{1, t}$ for the papers that appeared in mainly applied journals (for example, Energy Policy, Journal of International Money and Finance, Journal of Policy Modelling) is above $90 \%$. To the other extreme, the average $c_{2, t}$ is above $90 \%$ for the Journal of Econometrics and for Econometric Theory (but also for Econometrica, which hosted 24 citing papers). Other journals, such as Oxford Bulletin of Economics and Statistics, Journal of Applied Econometrics, Journal of Forecasting are more balanced, with an average $c_{2, t}$ around $50 \%$. We believe that the evidence in Table 3 supports the idea that classifying based on the title and the abstract is more accurate than classifying based on the publishing journal.

The time series $c_{1, t}$ and $c_{2, t}$ are illustrated in Figure 2. The plot shows some evidence of a "second wind" especially in the applied index $c_{1, t}$ but to some extent also in the methodological index $c_{2, t}$ : both series seem to have a peak around 1998, after which they start decreasing very slowly, but around 2004 the citations start increasing again, especially for the applied research index, whereas the references found in methodological papers remain rather steady. A possibility/conjecture is that the second wind was triggered by the 2003 Nobel Prize in Economics, which popularized the concept of cointegration in a wider
variety of scientific disciplines. The trajectory of $c_{1, t}$ resembles the cases of Friedrich Hayek, referred to in Bjork et al. (2014) as "bi-modal", whereas the trajectory of $c_{2, t}$ resembles more closely the cases of Kenneth Arrow and Milton Friedman, called "staying power" in Bjork et al. (2014). Boswijk et al. (2010) also claim the same with different wording: they report evidence of a "second life" for the famous Engle and Granger (1987) paper in Econometrica after the authors were awarded the Nobel prize in 2003, an event which is likely to have revamped the interest in the work of KJ and SJ as well.

Table 3. The top 20 journals supplying citations to S. Johansen and K. Juselius' works.

| Rank | Journal | $C_{T}$ | $\frac{C_{1, T}}{C_{T}}$ | $\frac{C_{2, T}}{\boldsymbol{C}_{T}}$ |
| ---: | :--- | ---: | ---: | ---: |
| 1 | APPLIED ECONOMICS | 539 | $92.0 \%$ | $8.0 \%$ |
| 2 | APPLIED ECONOMICS LETTERS | 254 | $89.2 \%$ | $10.8 \%$ |
| 3 | ENERGY ECONOMICS | 208 | $94.9 \%$ | $5.1 \%$ |
| 4 | ECONOMIC MODELLING | 204 | $92.7 \%$ | $7.3 \%$ |
| 5 | ENERGY POLICY | 172 | $95.1 \%$ | $4.9 \%$ |
| 6 | JOURNAL OF ECONOMETRICS | 154 | $2.9 \%$ | $97.1 \%$ |
| 7 | J. OF INTERNATIONAL MONEY \& FINANCE | 135 | $95.7 \%$ | $4.3 \%$ |
| 8 | JOURNAL OF POLICY MODELING | 121 | $96.4 \%$ | $3.7 \%$ |
| 9 | ECONOMICS LETTERS | 104 | $62.0 \%$ | $38.0 \%$ |
| 10 | JOURNAL OF MACROECONOMICS | 96 | $92.0 \%$ | $8.0 \%$ |
| 11 | OXFORD BULLETIN OF ECON. \& STAT. | 92 | $46.5 \%$ | $53.5 \%$ |
| 12 | ECONOMETRIC THEORY | 86 | $0.2 \%$ | $99.8 \%$ |
| 13 | EMPIRICAL ECONOMICS | 81 | $88.6 \%$ | $11.4 \%$ |
| 14 | JOURNAL OF FUTURES MARKETS | 70 | $96.1 \%$ | $3.9 \%$ |
| 15 | JOURNAL OF APPLIED ECONOMETRICS | 68 | $54.7 \%$ | $45.3 \%$ |
| 16 | MANCHESTER SCHOOL | 66 | $94.6 \%$ | $5.4 \%$ |
| 17 | ENERGY | 61 | $92.5 \%$ | $7.5 \%$ |
| 18 | JOURNAL OF BANKING \& FINANCE | 58 | $94.8 \%$ | $5.2 \%$ |
| 19 | JOURNAL OF BUSINESS \& ECON. STAT. | 55 | $38.4 \%$ | $61.6 \%$ |
| 20 | JOURNAL OF FORECASTING | 52 | $51.4 \%$ | $48.6 \%$ |
|  | All Journals | 6457 | $84.4 \%$ | $15.6 \%$ |



Figure 2. The composite citation indices $c_{1, t}$ (thick red, left scale) and $c_{2, t}$ (thin blue, right scale), quarterly data from 1989:1 to 2017:3.

## 3. The Bass Diffusion Model

The Bass diffusion model (Bass 1969) is widely used in many fields. Originally developed for marketing applications, the model has since been adopted also in other fields, such as the analysis of the diffusion of technological innovation (see Guseo and Guidolin 2008), bibliometric analysis (see Bjork et al. 2014) and epidemiology (see Eryarsoy et al. 2021).

The continuous time Bass model assumes a population of $m$ potential adopters. Let us define by $t>0$ the time of adoption of a randomly picked potential adopter: $t$ is therefore a random variable. Define by $f(t)$ the corresponding density, and by $F(t)=\int_{0}^{t} f(u) d u$ the cumulative density function, i.e., the probability that adoption occurs before $t$. Notice that the expected number of adopters at time $t$ is given by the following:

$$
\begin{equation*}
\bar{C}(t)=m F(t) \tag{3}
\end{equation*}
$$

and the corresponding "adoption intensity" is given by the following:

$$
\begin{equation*}
\bar{c}(t)=m f(t) \tag{4}
\end{equation*}
$$

Bass assumes that the hazard rate $\frac{f(t)}{1-F(t)}$ is a linear function of the expected number of previous adopters:

$$
\begin{equation*}
\frac{f(t)}{1-F(t)}=p+q F(t) \tag{5}
\end{equation*}
$$

where $q$ is defined as the "imitation parameter" (or internal influence, or word-of-mouth effect) since it represents the idea that some potential adopters (imitators) tend not to adopt initially, but are more likely to adopt when the innovation is widespread. Conversely, $p$ is defined as the "innovation parameter" (or external influence or advertising effect) since it represents the idea that some potential adopters (innovators) decide to adopt the innovation regardless of the level of diffusion. It is interesting to observe that when $q=0$, Equation (5) implies a constant hazard, and therefore the Bass model collapses into the exponential distribution. In other words, in the absence of imitators, the adoption peak, as in the exponential distribution, would occur at the beginning of the process. ${ }^{8}$

Using (3) and (4), the differential Equation (5) can be rewritten as follows:

$$
\begin{equation*}
\bar{c}(t)=m p+(q-p) \bar{C}(t)-\frac{q}{m} \bar{C}(t)^{2} . \tag{6}
\end{equation*}
$$

The solution to (6) with $\bar{C}(0)=0$ is the following:

$$
\begin{equation*}
\bar{C}(t)=m \frac{1-e^{-(p+q) t}}{1+\frac{q}{p} e^{-(p+q) t}}, \tag{7}
\end{equation*}
$$

so that

$$
\begin{equation*}
\bar{c}(t)=\frac{\partial \bar{C}(t)}{\partial t}=m \frac{p(p+q)^{2} e^{-(p+q) t}}{\left(p+q e^{-(p+q) t}\right)^{2}} \tag{8}
\end{equation*}
$$

Starting from the latter equation, one can easily find the timing of the adoption peak $t^{P}$ (i.e., the inflection point of the diffusion curve), the corresponding peak $\bar{c}^{P}$, and the level of adoption at the peak $\bar{C}^{P}$ :

$$
\begin{align*}
t^{P} & =\frac{1}{p+q} \ln \left(\frac{q}{p}\right)  \tag{9}\\
\bar{c}^{P} & =\bar{c}\left(t^{P}\right)=m \frac{(p+q)^{2}}{4 q}  \tag{10}\\
\bar{C}^{P} & =\bar{C}\left(t^{P}\right)=m \frac{q-p}{2 q} \tag{11}
\end{align*}
$$

Formula (9) shows that the location of the peak depends on the innovation parameter $p$ and the imitation parameter $q$ through the sum $(p+q)$ and the ratio $\frac{q}{p}$ : as clear in Formula (5), when either innovators or imitators or both are very active so that the sum
$(p+q)$ is large, then the hazard is large, which leads to a rapid exhaustion of the population at risk and therefore to an early peak.

### 3.1. Bass Discrete Time Model

A number of estimation procedures have been proposed to estimate the parameters $m$, $p$ and $q$ (see for example Satoh 2001). Bass (1969) suggested a simple estimation strategy based on Ordinary Least Squares (OLS) applied to a discretized version of (6) where essentially the expected adoption stock $\bar{C}(t)$ and the expected adoption flow $\bar{c}(t)$ are replaced by the observed counterpart $C_{t}$ and $c_{t}=C_{t}-C_{t-1}$, and an error term is added. This leads to the following:

$$
\begin{equation*}
c_{t}=m p+(q-p) C_{t-1}-\frac{q}{m} C_{t-1}^{2}+u_{t} . \tag{12}
\end{equation*}
$$

In the standard discrete time Bass model, $u_{t}$ is assumed to be $\operatorname{iidN}\left(0, \sigma^{2}\right)$, so that OLS is the natural candidate for estimation. To apply OLS, (12) is then reparameterized as follows:

$$
\begin{equation*}
c_{t}=\beta_{0}+\beta_{1} C_{t-1}+\beta_{2} C_{t-1}^{2}+u_{t} . \tag{13}
\end{equation*}
$$

The "reduced form" parameters $\beta=\left(\beta_{0}, \beta_{1}, \beta_{2}\right)^{\prime}$ are related to the "structural form" parameters $\boldsymbol{\theta}=(m, p, q)^{\prime}$ by the following:

$$
\begin{align*}
& \beta_{0}=m p \\
& \beta_{1}=q-p  \tag{14}\\
& \beta_{2}=-\frac{q}{m} .
\end{align*}
$$

and these relations can be inverted: ${ }^{9}$

$$
\begin{align*}
& m=\frac{-\beta_{1}-\sqrt{\beta_{1}^{2}-4 \beta_{0} \beta_{2}}}{2 \beta_{2}} \\
& p=\frac{\beta_{0}}{m}=\frac{-\beta_{1}+\sqrt{\beta_{1}^{2}-4 \beta_{0} \beta_{2}}}{2}  \tag{15}\\
& q=-m \beta_{2}=\frac{\beta_{1}+\sqrt{\beta_{1}^{2}-4 \beta_{0} \beta_{2}}}{2}
\end{align*}
$$

Assuming that $u_{t}$ is uncorrelated, homoskedastic and normal, ML estimates of the parameters vector, say $\hat{\boldsymbol{\beta}}$, can be obtained by OLS, and the corresponding variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_{\hat{\beta}}$ can be obtained as usual. ${ }^{10}$ Replacing $\hat{\boldsymbol{\beta}}$ in (15) instead of $\beta$ gives $\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}(\hat{\boldsymbol{\beta}})$. Defining by

$$
J_{\theta \cdot \beta}=\frac{\partial \theta}{\partial \beta^{\prime}}
$$

and using the delta method, the variance-covariance matrix associated to $\hat{\boldsymbol{\theta}}$ is given by the following:

$$
\begin{equation*}
\hat{\Sigma}_{\hat{\theta}}=\hat{J}_{\theta \cdot \beta} \hat{\Sigma}_{\hat{\beta}} \hat{J}_{\theta \cdot \beta}^{\prime}, \tag{16}
\end{equation*}
$$

where $\hat{J}_{\theta . \beta}$ is the estimated counterpart of $J_{\theta . \beta}$. Tedious computation shows the following:

$$
\boldsymbol{J}_{\boldsymbol{\theta} \cdot \boldsymbol{\beta}}=\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}^{\prime}}=\frac{1}{\delta_{\beta}}\left[\begin{array}{ccc}
1 & -\frac{\beta_{1}+\delta_{\beta}}{2 \beta_{2}} & \frac{\beta_{1}\left(\beta_{1}+\delta_{\beta}\right)-2 \beta_{0} \beta_{2}}{2 \beta_{2}^{2}}  \tag{17}\\
-\beta_{2} & \frac{\beta_{1}-\delta_{\beta}}{2} & -\beta_{0} \\
-\beta_{2} & \frac{\beta_{1}+\delta_{\beta}}{2} & -\beta_{0}
\end{array}\right]
$$

where $\delta_{\beta}=\sqrt{\beta_{1}^{2}-4 \beta_{0} \beta_{2}} \cdot \hat{J}_{\theta \cdot \beta}$ is therefore obtained by replacing $\hat{\beta}$ in (17).
We remark that when one considers $n$ "seemingly unrelated" equations such as (13), i.e.,

$$
\begin{equation*}
c_{i, t}=\beta_{0, i}+\beta_{1, i} C_{i, t-1}+\beta_{2, i} C_{i, t-1}^{2}+u_{i, t}, \quad i=1, \ldots, n, \tag{18}
\end{equation*}
$$

and the variance-covariance matrix of $\boldsymbol{u}_{t}=\left(u_{1, t}, \ldots, u_{n, t}\right)^{\prime}$, say $\Omega_{u}$, is not diagonal, then equation by equation OLS is no longer equivalent to ML. In this case, the likelihood can be maximized by iterated Seemingly Unrelated Regression Equations (SURE), obtaining $\hat{\boldsymbol{\beta}}_{i}=\left(\hat{\beta}_{0, i}, \hat{\beta}_{1, i}, \hat{\beta}_{2, i}\right)^{\prime}, i=1, \ldots, n, \hat{\Omega}_{u}$, and the variance-covariance matrix of $\hat{\boldsymbol{\beta}}=\left(\hat{\boldsymbol{\beta}}_{1}^{\prime}, \ldots, \hat{\boldsymbol{\beta}}_{n}^{\prime}\right)^{\prime}$, i.e.,

$$
\hat{\Sigma}_{\hat{\beta}}=\left[\begin{array}{ccc}
\hat{\Sigma}_{\hat{\beta}_{1}} & \cdots & \hat{\Sigma}_{\hat{\beta}_{n}, \hat{\beta}_{1}}^{\prime} \\
\vdots & \ddots & \vdots \\
\hat{\Sigma}_{\hat{\beta}_{n}, \hat{\beta}_{1}} & \cdots & \hat{\Sigma}_{\hat{\beta}_{n}}
\end{array}\right]
$$

Then, applying (15) and (16) to each pair $\left(\hat{\boldsymbol{\beta}}_{i}, \hat{\Sigma}_{\hat{\beta}_{i}}\right)$ it is easy to obtain the ML estimates of the structural parameters $\boldsymbol{\theta}_{i}=\left(m_{i}, p_{i}, q_{i}\right)^{\prime}$ and the associated variance-covariance matrices.

### 3.2. Boswijk and Franses Model

Boswijk and Franses (2005), henceforth BF, emphasize two major problems in the model (13):

- The assumption that $u_{t}$ is uncorrelated is at odds with the empirical evidence that deviations of the observed adoption path with respect to the ideal equilibrium path are persistent.
- The assumption that $u_{t}$ is homoskedastic is disputable since, at the beginning and at the end of the diffusion process, when $c_{t}$ is expected to be close to zero, the variance of $c_{t}$ is likely to be much smaller than around the peak; related to this, simulating (13) with an homoskedastic and Gaussian error is likely to produce negative values of $c_{t}$ in the initial and final phases of the diffusion.
To deal with the first problem, they propose the following alternative model: ${ }^{11}$

$$
\begin{equation*}
\Delta c_{t}=\alpha\left(c_{t-1}-m p-(q-p) C_{t-1}+\frac{q}{m} C_{t-1}^{2}\right)+u_{t} . \tag{19}
\end{equation*}
$$

To understand the relationship between (12) and (19) it is interesting to observe that, adding and subtracting $\alpha\left(-(q-p) C_{t-2}+\frac{q}{m} C_{t-2}^{2}\right)$ to the right hand side of (19), and rearranging, one obtains the following:

$$
\begin{equation*}
\Delta c_{t}=\alpha\left(c_{t-1}-m p-(q-p) C_{t-2}+\frac{q}{m} C_{t-2}^{2}\right)-\alpha\left((q-p) \Delta C_{t-1}-\frac{q}{m} \Delta C_{t-1}^{2}\right)+u_{t} \tag{20}
\end{equation*}
$$

To interpret (20), define the following:

$$
c_{t}^{*}=m p+(q-p) C_{t-1}-\frac{q}{m} C_{t-1}^{2},
$$

and notice that $c_{t}^{*}$ is the expected value or $c_{t}$ according to the Bass discrete time model (12). Using the notation $c_{t}^{*}$, (20) can be rewritten as the following:

$$
\begin{equation*}
\Delta c_{t}=\alpha\left(c_{t-1}-c_{t-1}^{*}\right)-\alpha \Delta c_{t}^{*}+u_{t} \tag{21}
\end{equation*}
$$

The parameter $\alpha$ is expected to be negative. The first term in (21), i.e., $\alpha\left(c_{t-1}-c_{t-1}^{*}\right)$, can be thought of as an Error-Correction Mechanism: if $c_{t-1}=c_{t-1}^{*}$, the ECM is ineffective; if instead $c_{t-1}>c_{t-1}^{*}$, then the ECM term partly corrects the disequilibrium by reducing $c_{t}$ with respect to $c_{t-1}$; conversely, if $c_{t-1}<c_{t-1}^{*}$, then, through the negative $\alpha, c_{t}$ will increase with respect to $c_{t-1}$. The second term in (21), i.e., $-\alpha \Delta c_{t}^{*}$, can be thought of as a "Target Seeking" Mechanism, which induces dynamics in $c_{t}$, even if $c_{t-1}=c_{t-1}^{*}$ and $u_{t}=0$ : in fact $\Delta c_{t}^{*}$ will be zero when $\Delta C_{t-1}$ (and therefore $\Delta C_{t-1}^{2}$ ) is zero, which happens when the target level $m$ is reached and therefore $C_{t-1}=C_{t-2}=m$. Another viewpoint on the BF model, seen as an $\operatorname{AR}(2)$ model for $C_{t}$ with state dependent parameters is given in Appendix $A$.

It is important to remark that the standard Bass model (12) is a special case of (19) with $\alpha=-1$, so that one can set up a test $H_{0}: \alpha=-1$ to decide which model is preferable. The interpretation of the parameters $m, p$ and $q$ is exactly the same in both models since (19) is a generalized version of the original Bass model, where the "adjustment intensity", instead of being fixed at -1 , is represented by the unrestricted parameter $\alpha$. For example, when $\alpha=-0.5$, only half of the disequilibrium observed at the end of a time unit is adjusted within the subsequent time unit: this gives rise to some persistence in the disequilibrium.

To deal with the second problem (heteroskedasticity), BF propose to model $u_{t}$ as the following:

$$
\begin{equation*}
u_{t}=c_{t-1}^{\phi} \varepsilon_{t}, \tag{22}
\end{equation*}
$$

where $\varepsilon_{t}$ is assumed to be uncorrelated and homoskedastic with variance $\sigma^{2}$, so that the variance of $u_{t}$ is assumed to be proportional to $c_{t-1}^{2 \phi}$; the authors do not consider $\phi$ as a parameter to be estimated, but they rather fix it heuristically to either $1 / 2$ or 1 , finding that $1 / 2$ is preferable in their application. In the application, we will use the residuals of the homoskedastic model to test for homoskedasticity vs. heteroskedasticity of the proposed type.

Model (19) can be reparameterized in different ways:

$$
\begin{align*}
\Delta c_{t} & =\alpha\left(c_{t-1}-\beta_{0}-\beta_{1} C_{t-1}-\beta_{2} C_{t-1}^{2}\right)+u_{t}  \tag{23}\\
& =\alpha c_{t-1}+\gamma_{0}+\gamma_{1} C_{t-1}+\gamma_{2} C_{t-1}^{2}+u_{t} . \tag{24}
\end{align*}
$$

The parametrization (24) is suited for estimation, either with OLS when $u_{t}$ is assumed to be uncorrelated and homoskedastic, or by WLS (dividing left and right by $c_{t-1}^{\phi}$ ), if $u_{t}$ is assumed to follow (22). Conversely, the parametrization (23) is useful because the parameters in $\beta=\left(\beta_{0}, \beta_{1}, \beta_{2}\right)^{\prime}$ are related to the parameters $\boldsymbol{\theta}=(m, p, q)^{\prime}$ as in (15): therefore if we obtain estimates of $\beta$ and $\Sigma_{\hat{\beta}}$, we can map them into estimates of $\boldsymbol{\theta}$ and $\Sigma_{\hat{\boldsymbol{\theta}}}$ using (15) and (16) directly. ${ }^{12}$

Let us define $\gamma^{\prime}=\left(\gamma_{0}, \gamma_{1}, \gamma_{2}\right)$. Assuming that $u_{t} \sim \operatorname{iidN}\left(0, \sigma^{2}\right)$, ML estimates of $\pi^{\prime}=\left(\alpha, \gamma^{\prime}\right)$ can be obtained by OLS in (24), obtaining $\hat{\alpha}, \hat{\gamma}$ and the corresponding variancecovariance matrix:

$$
\hat{\Sigma}_{\hat{\pi}}=\left[\begin{array}{cc}
\hat{\sigma}_{\alpha}^{2} & \hat{\Sigma}_{\hat{\gamma} \cdot \hat{\alpha}}^{\prime} \\
\hat{\Sigma}_{\hat{\gamma} \cdot \hat{\alpha}} & \hat{\Sigma}_{\hat{\gamma}}
\end{array}\right] .
$$

Notice that $\beta=-\frac{1}{\alpha} \gamma$; therefore, ML estimates of $\beta$ are given by the following:

$$
\begin{equation*}
\hat{\boldsymbol{\beta}}=-\frac{1}{\hat{\alpha}} \hat{\gamma} . \tag{25}
\end{equation*}
$$

We then obtain the following:

$$
J_{\beta . \pi}=\frac{\partial \beta}{\partial \pi^{\prime}}=\frac{1}{\alpha^{2}}\left[\begin{array}{cccc}
\gamma_{0} & -\alpha & 0 & 0 \\
\gamma_{1} & 0 & -\alpha & 0 \\
\gamma_{2} & 0 & 0 & -\alpha
\end{array}\right]=\frac{1}{\alpha^{2}}\left(\gamma,-\alpha \boldsymbol{I}_{3}\right),
$$

and, using the delta method, we have the following:

$$
\begin{equation*}
\hat{\Sigma}_{\hat{\beta}}=\hat{J}_{\beta \cdot \pi} \hat{\Sigma}_{\hat{\pi}} \hat{J}_{\beta \cdot \pi}^{\prime}=\frac{1}{\hat{\alpha}^{2}}\left(\frac{\hat{\gamma} \hat{\gamma}^{\prime}}{\hat{\alpha}^{2}} \hat{\sigma}_{\alpha}^{2}-\frac{\hat{\Sigma}_{\hat{\gamma} \cdot \hat{\alpha}} \hat{\gamma}^{\prime}+\hat{\gamma} \hat{\Sigma}_{\hat{\gamma} \cdot \hat{\alpha}}^{\prime}}{\hat{\alpha}}+\hat{\Sigma}_{\hat{\gamma}}\right) \tag{26}
\end{equation*}
$$

where $\hat{\boldsymbol{J}}_{\beta . \pi}$ is the estimated counterpart of $\boldsymbol{J}_{\beta . \pi}$. Starting from (25) and (26) one can obtain $\hat{\boldsymbol{\theta}}$ and $\hat{\Sigma}_{\hat{\theta}}$ using (15) and (16). In particular, replacing $\hat{\Sigma}_{\hat{\beta}}=\hat{J}_{\beta . \pi} \hat{\Sigma}_{\hat{\pi}} \hat{J}^{\prime}{ }_{\beta . \pi}$ in (16), one obtains the following:

$$
\hat{\Sigma}_{\hat{\theta}}=\hat{J}_{\theta \cdot \beta} \hat{\Sigma}_{\hat{\beta}} \hat{J}_{\theta \cdot \beta}^{\prime}=\hat{J}_{\theta \cdot \beta} \hat{J}_{\beta . \pi} \hat{\Sigma}_{\hat{\pi}} \hat{J}_{\beta . \pi^{\prime}} \hat{J}_{\theta \cdot \beta} \beta^{\prime}
$$

Additionally, in this case, when one considers $n$ "seemingly unrelated" equations such as (24), i.e.,

$$
\begin{equation*}
c_{i, t}=\alpha_{i} c_{i, t-1}+\gamma_{0, i}+\gamma_{1, i} C_{i, t-1}+\gamma_{2, i} C_{i, t-1}^{2}+u_{i, t}, \quad i=1, \ldots, n \tag{27}
\end{equation*}
$$

and the variance-covariance matrix of $\boldsymbol{u}_{t}=\left(u_{1, t}, \ldots, u_{n, t}\right)^{\prime}$, for example $\Omega_{u}$, is not diagonal, then equation by equation OLS is no longer equivalent to ML. In this case, the likelihood can be maximized by iterated SURE, obtaining $\hat{\pi}_{i}=\left(\alpha_{i}, \hat{\gamma}_{0, i}, \hat{\gamma}_{1, i}, \hat{\gamma}_{2, i}\right)^{\prime}, i=1, \ldots, n, \hat{\Omega}_{u}$, and the variance-covariance matrix of $\hat{\pi}=\left(\hat{\pi}_{1}^{\prime}, \ldots, \hat{\pi}_{n}^{\prime}\right)^{\prime}$, i.e.,

$$
\hat{\Sigma}_{\hat{\pi}}=\left[\begin{array}{ccc}
\hat{\Sigma}_{\hat{\pi}_{1}} & \cdots & \hat{\Sigma}_{\hat{\pi}_{n}, \hat{\pi}_{1}}^{\prime} \\
\vdots & \ddots & \vdots \\
\hat{\Sigma}_{\hat{\pi}_{n}, \hat{\pi}_{1}} & \cdots & \hat{\Sigma}_{\hat{\pi}_{n}}
\end{array}\right]
$$

Then, starting from each pair $\left(\hat{\boldsymbol{\pi}}_{i}, \hat{\Sigma}_{\hat{\pi}_{i}}\right)$ one can obtain the ML estimates of the structural parameters $\boldsymbol{\theta}_{i}=\left(m_{i}, p_{i}, q_{i}\right)^{\prime}$ and the associated variance-covariance matrices as illustrated above.

As for the asymptotic properties of ML estimates of the structural parameters, Boswijk and Franses (2005) prove that $\hat{m}$ is consistent in $T$ (as the time span increases $C_{T}$ ideally coincides with $m$ ), whereas $\hat{p}$ and $\hat{q}$ are not; moreover, they show that the asymptotic distribution cannot be proved to be normal. However, they demonstrate with an extensive simulation that when the frequency is allowed to go to infinity along with the time span, then $\hat{m}, \hat{p}$ and $\hat{q}$ are essentially unbiased and asymptotically normal; they also show that this is approximately valid, even with a fixed time span, at least if it includes the inflection point $\frac{\ln q-\ln p}{p+q}$. In other words, if the observed time span includes the inflection point and the sampling frequency is reasonably high, their results suggest that using the standard normal and the $\chi^{2}$ for making inference on the parameters is a reasonable approximation.

### 3.3. Boswijk et al. Multivariate Model

Boswijk et al. (2009), henceforth BFF, propose a multivariate generalization of (19). The BFF model is made up of $n$ equations, and can be written as follows:

$$
\begin{equation*}
\Delta c_{i, t}=\sum_{j=1}^{n} \alpha_{i j}\left(c_{j, t-1}-\left(p_{j}+q_{j}\right) C_{j, t-1}+\frac{q_{j}}{m_{j}} C_{j, t-1}^{2}-p_{j} m_{j}\right)+u_{i, t} \quad i=1, \ldots, n, \tag{28}
\end{equation*}
$$

where, in a simplified homoskedastic version of the model, we might assume that $u_{t}=$ $\left[u_{1, t}, \ldots, u_{n, t}\right] \sim$ iidN $N_{n}(\mathbf{0}, \boldsymbol{\Omega}) .{ }^{13}$ Along the lines of the BF model, (28) may be reparametrized as follows:

$$
\begin{equation*}
\Delta c_{i, t}=\sum_{j=1}^{n} \alpha_{i j}\left(\beta_{j}^{\prime} X_{j, t-1}-\beta_{0 j}\right)+u_{i t}, \quad i=1, \ldots, n \tag{29}
\end{equation*}
$$

with

$$
\boldsymbol{X}_{j t}^{\prime}=\left[c_{j, t}, C_{j, t}, C_{j, t}^{2}\right], \quad \boldsymbol{\beta}_{j}^{\prime}=\left[1,-\beta_{1 j},-\beta_{2 j}\right]
$$

or, more compactly,

$$
\begin{equation*}
\boldsymbol{Y}_{t}=\alpha \boldsymbol{\beta}^{\prime} \boldsymbol{X}_{t-1}+\boldsymbol{u}_{t} \tag{30}
\end{equation*}
$$

where

$$
\begin{aligned}
& \underset{n \times 1}{\boldsymbol{Y}_{t}}=\left[\begin{array}{c}
\Delta c_{1, t} \\
\vdots \\
\Delta c_{n, t}
\end{array}\right], \underset{(3 n+1) \times 1}{\boldsymbol{X}_{t}}=\left[\begin{array}{c}
X_{1, t} \\
\vdots \\
X_{n, t} \\
1
\end{array}\right], \underset{n \times 1}{\boldsymbol{u}_{t}}=\left[\begin{array}{c}
u_{1, t} \\
\vdots \\
u_{n, t}
\end{array}\right], \\
& \underset{n \times n}{\boldsymbol{\alpha}}=\left[\begin{array}{ccc}
\alpha_{11} & \cdots & \alpha_{1 n} \\
\vdots & \ddots & \vdots \\
\alpha_{n 1} & \cdots & \alpha_{n n}
\end{array}\right], \underset{(3 n+1) \times n}{\beta}=\left[\begin{array}{ccc}
\beta_{1} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & \boldsymbol{\beta}_{n} \\
\beta_{01} & \cdots & \beta_{0 n}
\end{array}\right] .
\end{aligned}
$$

Since this paper's main goal is to celebrate Søren Johansen and Katarina Juselius, it is nice to remark that, apart from the exclusion restrictions in $\beta$, and the fact that the rank of $\alpha \beta^{\prime}$ is actually full, (30) has the mathematical form of the "reduced rank regression" popularized by Søren and Katarina; therefore, in estimating and interpreting the model, we can benefit directly from the results inspired by their work, in particular Hansen (2003). Notice that the (exclusion) restrictions on the matrix $\beta$ can be written as the following:

$$
\begin{equation*}
\operatorname{vec}(\beta)=\boldsymbol{H}_{\beta} \varphi_{\beta}+\boldsymbol{h}_{\beta} \tag{31}
\end{equation*}
$$

for suitable restriction matrices $H_{\beta}$ and $h_{\beta} .{ }^{14}$ It might be also interesting to consider restrictions on $\alpha$ of the following type:

$$
\begin{equation*}
\operatorname{vec}(\alpha)=H_{\alpha} \varphi_{\alpha^{\prime}} \tag{32}
\end{equation*}
$$

for example to test the hypothesis that the matrix $\alpha$ is diagonal, under which (28) would collapse into $n$ "seemingly unrelated" BF equations such as (19). ${ }^{15}$ Of course, when $\alpha$ is unrestricted, we have that $H_{\alpha}=I_{n^{2}}$ and $\varphi_{\alpha}=\operatorname{vec}(\boldsymbol{\alpha})$.

Assuming that $\boldsymbol{u}_{t} \sim \operatorname{iid} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$, the log-likelihood function is given by $\ell(\boldsymbol{\varphi}, \boldsymbol{\Omega})=$ $-\frac{T}{2}\left[n \ln (2 \pi)+\ln |\boldsymbol{\Omega}|+\operatorname{tr}\left(\boldsymbol{\Omega}^{-1} \boldsymbol{M}_{u u}\right)\right]$, where $\boldsymbol{M}_{u u}=T^{-1} \sum_{t=1}^{T} \boldsymbol{u}_{t} \boldsymbol{u}_{t}^{\prime}$. Since the log-likelihood score is bi-linear in the parameters $\alpha$ and $\beta$, one can employ the generalized reduced rank regression algorithm proposed by Hansen (2003) for likelihood maximization of I(1) VAR models under linear restrictions. This provides maximum likelihood estimates of the parameters $\varphi=\left[\varphi_{\alpha}^{\prime}, \varphi_{\beta}^{\prime}\right]^{\prime}$ and $\Omega$, for example, $\hat{\boldsymbol{\varphi}}$ and $\hat{\Omega} .{ }^{16}$

To work out the variance-covariance matrix associated to $\hat{\varphi}$, notice that the model (30) under the restriction (31) and (32) is a sub-model of the following regression model:

$$
\boldsymbol{Y}_{t}=\boldsymbol{\Pi} \boldsymbol{X}_{t-1}+\boldsymbol{u}_{t}
$$

where $\Pi=\alpha \beta^{\prime}=\Pi(\varphi)$ is a smooth function of the vector of the parameters in $\varphi$. The second derivatives of the log-likelihood with respect to vec $(\boldsymbol{\Pi})$ are given by $-T \boldsymbol{M}_{X X} \otimes \boldsymbol{\Omega}^{-1}$, see e.g., Johansen (2006, Equation (13)), where $\boldsymbol{M}_{X X}=T^{-1} \sum_{t=1}^{T} \boldsymbol{X}_{t-1} \boldsymbol{X}_{t-1}^{\prime}$. Because the
parameters in $\varphi$ and in $\Omega$ are asymptotically independent, one finds that the Hessian with respect to $\varphi$ equals the following:

$$
\begin{equation*}
\mathcal{H}_{\varphi}=\frac{\partial^{2} \ell(\varphi, \Omega)}{\partial \varphi \partial \boldsymbol{\varphi}^{\prime}}=-T J_{\Pi . \varphi}^{\prime}\left(\boldsymbol{M}_{X X} \otimes \Omega^{-1}\right) J_{\Pi . \varphi^{\prime}} \tag{33}
\end{equation*}
$$

where

$$
J_{\Pi . \varphi}=\frac{\partial \operatorname{vec} \Pi(\varphi)}{\partial \varphi^{\prime}}
$$

In order to describe $J_{\Pi . \varphi}$ in more detail observe that, in the present case, one has $\operatorname{vec} \boldsymbol{\Pi}(\boldsymbol{\varphi})=\operatorname{vec} \boldsymbol{\alpha} \boldsymbol{\beta}^{\prime}=\left(\boldsymbol{\beta} \otimes \boldsymbol{I}_{n}\right) \operatorname{vec} \boldsymbol{\alpha}=\left(\boldsymbol{I}_{(3 n+1)} \otimes \boldsymbol{\alpha}\right) \operatorname{vec} \boldsymbol{\beta}^{\prime}$. Therefore, using (31) and (32), one finds the following:

$$
\begin{aligned}
\partial \operatorname{vec}\left(\boldsymbol{\alpha} \boldsymbol{\beta}^{\prime}\right) / \partial \boldsymbol{\varphi}_{\alpha}^{\prime} & =\left(\boldsymbol{\beta} \otimes \boldsymbol{I}_{n}\right) \boldsymbol{H}_{\boldsymbol{\alpha}} \\
\partial \operatorname{vec}\left(\boldsymbol{\alpha} \boldsymbol{\beta}^{\prime}\right) / \partial \boldsymbol{\varphi}_{\beta}^{\prime} & =\left(\boldsymbol{I}_{(3 n+1)} \otimes \boldsymbol{\alpha}\right) \mathcal{K}_{(3 n+1), n} \boldsymbol{H}_{\boldsymbol{\beta}}
\end{aligned}
$$

where $\mathcal{K}_{m n}$ is a commutation matrix, which satisfies $\mathcal{K}_{m n} \operatorname{vec}(\boldsymbol{M})=\operatorname{vec}\left(\boldsymbol{M}^{\prime}\right)$ when $\boldsymbol{M}$ is $m \times n$. Therefore

$$
\boldsymbol{J}_{\Pi . \varphi}=\operatorname{blkdiag}\left(\left(\boldsymbol{\beta} \otimes \boldsymbol{I}_{n}\right) \boldsymbol{H}_{\alpha}\left(\boldsymbol{I}_{(3 n+1)} \otimes \boldsymbol{\alpha}\right) \mathcal{K}_{(3 n+1), n} \boldsymbol{H}_{\boldsymbol{\beta}}\right) .
$$

The variance-covariance matrix of $\hat{\varphi}$ can be then estimated by the following:

$$
\hat{\Sigma}_{\hat{\varphi}}=-\hat{\mathcal{H}}_{\varphi}^{-1}
$$

where $\hat{\mathcal{H}}_{\varphi}$ is obtained by plugging the ML estimates $\hat{\varphi}$ and $\hat{\Omega}$ instead of $\boldsymbol{\varphi}$ and in $\boldsymbol{\Omega}$ in (33).

## 4. Results

Our statistical analysis is based on two equations, headed to $c_{1, t}$ and $c_{2, t}$, respectively (see Section 2 for a definition of the indices). In this section, we will first discuss the estimates of the reduced form models and then the corresponding estimates of the structural form models.

### 4.1. Analysis of the Reduced Form-Comparing Bass, BF, BFF

As illustrated in the previous section, the two univariate Bass Equation (18) may be seen as a restricted version of the bivariate BFF model (29), with four restrictions: $\alpha_{11}=\alpha_{22}=-1$ and $\alpha_{12}=\alpha_{21}=0$. Similarly, the univariate dynamic BF model (27) may also be seen as a restricted version of the bivariate BFF model (29), with only two restrictions: $\alpha_{12}=\alpha_{21}=0$. In all cases, assuming that the errors in the two equations are simultaneously correlated, i.e.,

$$
\mathbf{\Omega}_{u}=\left[\begin{array}{cc}
\sigma_{1}^{2} & \rho \sigma_{1} \sigma_{2} \\
\rho \sigma_{1} \sigma_{2} & \sigma_{2}^{2}
\end{array}\right],
$$

efficient estimates of all models may be obtained by maximum likelihood, using the Hansen (2003) algorithm as illustrated in Section 3.3. ${ }^{17}$ The results are shown in Table 4.

Table 4. ML estimates of the reduced form parameters.

|  |  | Model (18) |  | Model (27) |  | Model (29) |  |
| :--- | :--- | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Estimate | t-Ratio | Estimate | t-Ratio | Estimate | t-Ratio |
| APP | $\hat{\alpha}_{1,1}$ | -1 |  | -0.435 | -5.74 | -0.508 | -6.26 |
|  | $\hat{\alpha}_{1,2}$ | 0 |  | 0 |  | 0.715 | 2.81 |
|  | $\hat{\beta}_{0,1}$ | 17.55 | 9.00 | 19.40 | 5.28 | 19.81 | 4.78 |
|  | $\hat{\beta}_{1,1}$ | 0.0196 | 9.62 | 0.0183 | 4.76 | 0.0190 | 4.61 |
|  | $\hat{\beta}_{2,1}$ | $-1.33 \times 10^{-6}$ | -3.22 | $-1.054 \times 10^{-6}$ | -1.35 | $-1.32 \times 10^{-6}$ | -1.60 |
|  | $\hat{\sigma}_{1}$ | 10.40 |  | 8.42 |  | 8.15 |  |
|  | $\hat{\alpha}_{2,1}$ | 0 |  | 0 |  | 0.0635 | 2.34 |
|  | $\hat{\alpha}_{2,2}$ | -1 |  | -0.492 | -6.28 | -0.547 | -6.44 |
|  | $\hat{\beta}_{0,2}$ | 5.90 | 8.28 | 6.27 | 5.05 | 6.74 | 4.65 |
|  | $\hat{\beta}_{1,2}$ | 0.0205 | 5.77 | 0.0199 | 3.21 | 0.0187 | 2.73 |
|  | $\hat{\beta}_{2,2}$ | $-2.06 \times 10^{-5}$ | -5.71 | $-2.07 \times 10^{-5}$ | -3.29 | $-1.97 \times 10^{-5}$ | -2.86 |
|  | $\hat{\sigma}_{2}$ | 3.30 |  | 2.79 |  | 2.72 |  |

Table 5 reports some misspecification tests based on the residuals illustrated in Figure 3.

The two rows, headed AC, in the table report the results of tests for auto correlation of the residuals of the applied and methodological equation, respectively. Specifically, we tested for serial correlation up to $k=20$ lags using the Ljung-Box Q-statistic, whose null hypothesis is that the errors are uncorrelated. ${ }^{18}$ The p-value is zero for the standard Bass model (18): therefore, residuals serial correlation is a major problem for that model. Conversely, in models (27) and (29), the white noise assumption is not rejected for the methodological equation, while for the applied equation, there is a clear improvement over model (18), but some autocorrelation seems to remain for both models, which suggests to invest more on the dynamic specification, which is left for further research.

The four rows headed HSK in the table report the results of two different types of BreuschPagan tests for heteroskedasticity for the applied and methodological equation, respectively In all cases, the null hypothesis is that the errors are homoskedastic, but we introduced two different alternatives. In fact, as seen in Equation (22), Boswijk and Franses (2005) suggest that the standard deviation should be proportional to $c_{i, t-1}^{\phi}$ (with $\phi=1 / 2$ or $\phi=1$ ); therefore, we introduced the constant and $c_{i, t-1}^{2 \phi}$ in the auxiliary regression, with two alternative values for $\phi$. For model (18), the null is rejected in most cases. ${ }^{19}$ Conversely, in spite of the very convincing argument supporting heteroskedasticity made by the cited authors, we did not find statistically significant evidence in this sense for this data set in (27) and (29); therefore, for the analysis in this paper, we did not consider the heteroskedastic versions of BF and BFF models.

The log-likelihood increases by 33.75 from model (18) to model (27): the LR test is therefore $\chi_{2}^{2}=67.50$, and the p -value is essentially zero. According to this result, the standard Bass model seems unable to capture the persistent swings clearly visible in Figure 2 and in the first plot of Figure 3: notice in fact that both parameters $\alpha_{11}$ and $\alpha_{22}$ estimated in (27) are approximately -0.5 and statistically different from -1 , which implies that only half of the distance from the ideal Bass path is corrected within one quarter, giving rise to persistent disequilibria. However, even model (27) is not satisfactory: in fact, the $\log$-likelihood of model (29) is significantly higher (the LR test is $\chi_{2}^{2}=14.6, p$-value 0.00111 ). This result is interesting since it suggests the existence of Granger causality running from the methodological research to applied research and/or vice-versa.

Table 5. Residual based tests for autocorrelation and heteroskedasticity. AC: Ljung-Box Q-statistics up to $k=20$ lags (see footnote 25 for $p$ ). HSK: Breusch-Pagan test, including the constant and $c_{1, t-1}^{2 \phi}$ (or $c_{2, t-1}^{2 \phi}$ ) in the regression where the dependent variable is $\frac{\hat{u}_{1, t}^{2}}{\sigma_{1}^{2}}$ (or $\frac{\hat{u}_{2, t}^{2}}{\hat{\sigma}_{2}^{2}}$.

|  |  | Model (18) |  | Model (27) |  | Model (29) |  |
| :--- | :--- | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Test | $p$-Value | Test | $p$-Value | Test | $p$-Value |
| APP | AC $\chi_{k-p}^{2}$ | $H S K(\phi=1 / 2) \chi_{1}^{2}$ | 3.96 | 0.047 | 0.88 | 0.348 | 2.79 |
|  | $H S K(\phi=1) \chi_{1}^{2}$ | 3.86 | 0.047 | 0.33 | 0.566 | 1.27 | 0.095 |
|  | AC $\chi_{k-p}^{2}$ | 101.1 | 0.000 | 27.48 | 0.051 | 26.31 | 0.069 |
|  | $H S K(\phi=1 / 2) \chi_{1}^{2}$ | 1.41 | 0.235 | 1.77 | 0.184 | 2.96 | 0.085 |
|  | $H S K(\phi=1) \chi_{1}^{2}$ | 6.30 | 0.012 | 0.91 | 0.340 | 1.46 | 0.226 |



Figure 3. Residuals of different models: Bass = model (18), $\mathrm{BF}=$ model (27), BFF $=$ model (29). Thick red line $=$ "Applied" (left scale). Thin blue line $=$ "Methodological" (right scale).

To shed some light on this, we observe that the estimates of $\alpha_{12}$ and $\alpha_{21}$ in (29) are both positive and statistically significant, suggesting that an increase in the methodological research leads to expect more applications in the future, and that an increase in applications stimulates further methodological research, with a continuous dialogue between the economic problems and econometric methods, which is exactly in the spirit of KJ and SJ's main message to the profession.

To provide a visual illustration of the relevance of the dynamic interaction between methodological and applied research in this field, we carry on a simulation exercise, similar in spirit to impulse response analysis. Impulse response functions, being the reactions of the variables to shocks entering the system, are useful for studying the interactions between variables in a vector autoregressive model (Lütkepohl 2016). In a more general non-linear setting, Potter (2000) and Koop et al. (1996) remark that nonlinear models produce impulse responses that are history- and shock-dependent; to overcome this problem, they introduce the notion of "generalized impulse response functions", based on a stochastic simulation, which can be applied in both the linear and non-linear case. We considered this tool, but since the non-linearity is relatively mild in our case, we opted for a tailored solution that is closer to the traditional deterministic impulse response analysis.

We initialize $C_{i, 0}=c_{i, 0}=0,{ }^{20}$ and then we compute two alternative trajectories for $c_{i, t}$ based on the estimated counterpart of (29). In the first dynamic simulation, $u_{i, t}$ is set to zero for all $i$ and $t$ : this leads to the "unshocked" paths $c_{i, t}^{U}$, corresponding to the deterministic trajectory that would take place in the absence on any innovation, starting from the assumed initial conditions. In the second dynamic simulation, we set $u_{j, 1}=\hat{\sigma}_{j}$ in the $j$-th equation, whereas all other innovations (different equations and/or different times) are set to zero so that the impulse corresponds to one standard deviation in just one of the equations; ${ }^{21}$ this exercise leads to the shocked paths $c_{j, i, t}^{S}$, where the first subscript, $j$, indicates which equation has been shocked. The standardized response of the $i$-th equation to an impulse on the $j$-th equation are then given by the difference of the two trajectories, standardized by the standard deviation of the output variable as follows:

$$
I R_{j, i, t}=\frac{\left(c_{j, i, t}^{S}-c_{i, t}^{U}\right)}{\hat{\sigma}_{i}} \quad j=1, \ldots, n ; \quad i=1, \ldots, n ; \quad t=1, \ldots
$$

The IRs therefore isolate that part of the trajectory $c_{j, i, t}^{S}$, which can be attributed to the shock. The first 20 IRs are illustrated in Figure 4. Notice that, by construction, $I R_{j, i, 1}$ is equal to 1 for $j=i, 0$ otherwise. According to Figure 4, in the short run, the (standardized) response of the methodological literature to a (standardized) impulse in the applied literature appears qualitatively very similar to the (standardized) response of the applied literature to a (standardized) impulse in the methodological literature.


Figure 4. Standardized impulse responses based on model (29). Initialization: $C_{i, 0}=c_{i, 0}=0, i=1,2$.
Some more insight on the relationship between methodological and applied research can be obtained by analyzing the cumulative IRs. It is important to remark that, given the mathematical nature of the model, the shocks do not have permanent effects. In fact, as $t$ goes to infinity, the cumulative citations $C_{i, t}$ will eventually reach the saturation point $m_{i}$ irrespective of the initial conditions and/or the shocks they undergo: this implies that $\sum_{t=1}^{\infty} c_{i, t}^{U}=\sum_{t=1}^{\infty} c_{j, i, t}^{S}=m_{i}-C_{i, 0}$ for any $i, j$ and $C_{i, 0}$, and therefore $\sum_{t=1}^{\infty} I R_{j, i, t}=0$. As a consequence, although the first IRs illustrated in Figure 4 are positive, at some point they turn negative (although with a very small magnitude) so that, in the limit, the cumulative sum is zero. This behavior is better illustrated through the cumulative IRs, illustrated in Figure 5, for a much longer period (500 quarters).


Figure 5. Cumulative standardized impulse responses based on model (29). Initialization: $C_{i, 0}=c_{i, 0}=0, i=1,2$.

Figure 5 shows that an impulse equal to $\hat{\sigma}_{1}$ (i.e., 8.15 papers) in the applied literature is strongly "self exciting", giving rise to a very long sequence of positive IRs in the applied literature itself, adding up to $10.5 \hat{\sigma}_{1}$ (about 85 papers) in the subsequent 150 quarters (almost 40 years), before it starts fading away. Conversely, the cumulative impact on the methodological literature of the same impulse is shorter living, and way less relevant (1.2 $\hat{\sigma}_{2}$, i.e., about 3 papers). On the other hand, an impulse equal to $\hat{\sigma}_{2}$ (i.e., 2.72 papers) in the methodological literature is not so "self-exciting" (the peak of the cumulative IRs is only $3.6 \hat{\sigma}_{2}$ - 10 papers-about 50 quarters after the impulse), whereas the cumulative impact on the applied literature seems very important (the peak is equal to $4.4 \hat{\sigma}_{1}-35$ papers-about 150 quarters after the impulse). This evidence seems to suggest that, although in the short run, the cross fertilization is rather balanced, in the long run, the methodological literature triggers the applications more than the other way around. ${ }^{22}$

Actually, the extremely long sequence of positive IR's, well beyond the observed period of 115 quarters, casts some doubt on the validity of the implicit assumption that the impulses do not have a permanent effect. We think that a hint for future research arising from the current study is to develop an alternative model where the saturation point $m_{i}$ is not already set at the beginning of the process, but it is to some extent "path dependent". In fact, if an idea appears more successful than what was initially assumed (i.e., we observe some unexpected citations), we should reconsider the expected total number of citations in the long run, leading to an upward revision. Conversely, when an idea is suddenly abandoned, possibly in favor of an alternative paradigm (i.e., we observe an unexpected reduction in the number of citations), we should reasonably revise downwards the expected total number of citations in the long run.

### 4.2. Analysis of the Structural Form

Table 6 reports the "structural" parameters $m, p$ and $q$ in the models (12), (19) and (28), which are based on the ML estimates of (18), (27) and (29), respectively. The associated standard errors are computed using the delta method, as illustrated in Sections 3.1-3.3. It is important to remark that the standard errors reported for models (12) are not reliable: they appear to be much lower than in the other two models, but the assumptions for applying ML—in particular, the absence of serial correlation—are clearly invalid for that model as illustrated in Table 5.

Table 6. Estimates of the structural parameters of the Bass, BF and BFF models. The standard errors of the $\hat{t}^{P}$ 's measured in quarters.


The timing of the citations peaks $\hat{t}_{1}^{P}$ and $\hat{t}_{2}^{P}$, the associated peaks $\hat{\bar{c}}_{1}^{P}$ and $\widehat{\bar{c}}_{2}^{P}$, and the corresponding cumulative number of citations at the peak $\widehat{\widetilde{C}}_{1}^{P}$ and $\widehat{\bar{C}}_{2}^{P}$, are obtained by plugging the estimated structural parameters in (9)-(11), and the associated standard errors are computed using the delta method. ${ }^{23}$

According to the evidence provided in Table 6, the estimates of the structural parameters are rather robust to the model used. Our comments are focused on the results based on model (28), which is statistically preferable.

It is interesting, and not surprising, that the "innovation parameter" $p$ is much higher for the methodological literature, whereas the "imitation parameter" is quite similar in the two strands of the literature: this makes imitation relatively more important than innovation in the applied literature. As for the citation peaks, it seems that the peak in the methodological literature (11 papers per quarter) was reached in 2001, whereas the peak in the applied literature (88 papers per quarter) is expected in 2020, 12 quarters after the end of the estimation sample (although the associated standard error is extremely large- 32 quarters). Based on the discussion of the properties of the estimates provided in Boswijk and Franses (2005), the estimates of the methodological equation should, therefore, be regarded as more reliable since the inflection point of the diffusion curve appears to be within the sample; this is less so for the estimates of the applied equation, where the estimated inflection point is outside the sample (of course we do not know the "true" inflection point). Not surprisingly, the standard error of $\hat{m}_{1}$ is quite large (the coefficient of variation $\hat{\sigma}_{\hat{m}_{1}} / \hat{m}_{1}$ is about $42 \%$ ), while the standard error of $\hat{m}_{2}$ is much smaller (the coefficient of variation $\hat{\sigma}_{\hat{m}_{2}} / \hat{m}_{2}$ is less than $10 \%$ ).

Figure 6 illustrates the observed time series along with the estimated unconditional expectation obtained by plugging the estimated structural parameters in Equation (8). Strictly speaking, since at the end of the sample (2017:Q3) we observe $C_{1, T}=5445.2$ and $C_{2, T}=1011.8$, our point estimates would ideally imply that we should expect $\hat{m}_{1}-C_{1, T}=9905$ applied WoS papers and $\hat{m}_{2}-C_{2, T}=216$ methodological WoS papers citing KJ and SJ in the future. We think that this interpretation is hazardous, to say the least. It is worth observing that the estimates of the structural parameters, especially $m_{1}$ and $m_{2}$, are very unstable as observed among others in Chandrasekaran and Tellis (2018), and they mainly seem to represent the history of the process in a descriptive sense rather than being a reliable forecasting tool in an inferential sense. For example, if we re-estimate the parameters based on the sub sample 1989:Q1-2005:Q4, so that the end of the sample occurs right before the "second wind" clearly visible in the plot, we would obtain $\hat{m}_{1}=2710.6$
$\left(\hat{\sigma}_{\hat{m}_{1}}=142.5\right), \hat{m}_{2}=684.3\left(\hat{\sigma}_{\hat{m}_{2}}=34.7\right), \hat{t}_{1}^{P}=1999: \mathrm{Q} 1$ and $\hat{t}_{2}^{P}=1997: Q 4 .{ }^{24}$ Therefore, estimating the same model 15 years ago, one would be convinced of the following: (i) that the citations peak was already reached several years before (and then the estimates would be regarded as reliable); (ii) that the potential for this literature was about one fifth of what it appears now; and (iii) that by 2020, the interest in KJ and SJ work will have disappeared (see Figure 7). However, as this Special Issue confirms, no prediction could have proved more wrong!


Figure 6. Observed time series along with the estimated unconditional expectation based on Equation (8). Left-applied index; right-methodological index.


Figure 7. Forecasting fallacy: observed time series along with the estimated unconditional expectation based on (8), parameters re-estimated based on the trimmed sample 1898:Q1-2005:Q4. Left-applied index; right-methodological index.

## 5. Conclusions and Suggestions for Further Research

Our main purpose in writing this paper was to contribute to the Festschrift in honor of Katarina Juselius and Søren Johansen as a sign of gratitude for their being for us a constant source of inspiration. We tried to find a way to show how profoundly they contributed to the development of economic ideas, emphasizing one key aspect of their approach, namely, the dialogue between empirical economics and econometric methodology. To this aim, we have proposed an operational way to disentangle, as much as possible, their contribution to applied and methodological econometric research, through the development of two indices based on the Web of Science database. We hope that this can also be a contribution
to bibliometric studies since a similar approach to assess in a quantitative way the impact of new ideas on methodological and applied research, and on the interaction between them, can be used for other areas. Ideally, similar analyses might be employed to investigate even more general epistemological issues, such as the relationship between theoretical and empirical research.

We think that the data we describe in Section 2 are very interesting per se. They show that KJ and SJ's influence on the literature is extremely important: their top 10 papers sum up to about 10,500 WoS citations (more than 50,000 in GS) from about 6500 citing papers, an average of more than 200 papers per year. Based on our indicators, $85 \%$ of the citing papers are essentially applied, whereas $15 \%$ are methodological: we do not have a benchmark for comparison, but we have the impression that the share of methodology is somewhat larger than in the econometric literature in general. As of 2017, the number of applied citing papers per quarter had not yet reached the peak (although a "false peak" seems to have occurred around 2000); conversely, the peak in the methodological literature seems to have been reached around 2001, although the shape of the trajectory is very flat after the peak, similar to what Bjork et al. (2014) has identified in a minority of Nobel prize winners and defined as "staying power".

To model the data, we resorted to an innovative dynamic multivariate versionproposed in Boswijk et al. (2009)—of the well-known Bass (1969) model. It was a pleasure for us to observe and emphasize that this model resembles so closely the Vector ECM model popularized by KJ and SJ; in particular, the bilinear nature of the model allows to use the Hansen (2003) algorithm to maximize the likelihood, which generalizes Johansen's ML algorithm, adapting it to a rather general class of restrictions, which includes our case.

The estimated model conveys very interesting information. As seen in Formula (9), the location of the citations peaks depends on the relative importance of the "innovation parameter" $p$ and the "imitation parameter" $q$. Our estimates suggest that the different location of the peaks might be explained by the higher value of the parameter $q_{2}$ with respect to $q_{1}$, whereas $p_{1}$ and $p_{2}$ are quite similar: using the standard terminology in the Bass model literature, the difference in the parameters suggests that the methodological literature is mainly driven by "innovators", whereas "imitators" are relatively more important in the applied literature.

Another interesting finding is that, in the literature referring to KJ and SJ , the "crossfertilization" between methodological and applied research is statistically significant and bi-directional (although possibly more effective from methodology to applications than the other way round). According to our impulse response analysis, rounding our figures, 8 unexpected applied papers in one quarter lead to predict that 3 methodological papers will follow, whereas 3 unexpected methodological papers lead to predict that 40 applied papers will follow (this is not so unbalanced as it seems at first sight since the scale of the two strands of literature is different). These results testify that one of the most important messages that Katarina Juselius and Søren Johansen have emphasized in their writingsi.e., that the applications should pose challenging problems to the methodology and that the methodology should sharpen the ability of applied researchers to ask meaningful questions to the data-has become a common heritage in this literature.

As for the estimated dimension of KJ and SJ influence, as measured by the parameters $m_{1}$ and $m_{2}$ (often called "saturation point" or "ceiling"), a word of caution is in order. Our estimates, $\hat{m}_{1}=15,351$ and $\hat{m}_{2}=1228$, imply that we should expect about 10,000 applied WoS papers and 200 methodological WoS papers citing KJ and SJ in the future. We do not consider these figures very reliable. Indeed, early in the literature, it was pointed out by Heeler and Hustad (1980) and others (e.g., Hyman 1988) that the predicting ability of the Bass model depends on the generation of accurate estimates of $m$. Srinivasan and Mason (1986) report problems with convergence when the data set does not contain the peak time period (i.e., the inflection point of the curve). The parameter $m$ is, again, under attack in Van den Bulte and Lilien (1997): there is evidence of downward bias in the estimation of the saturation point. Finally, in their review article Chandrasekaran
and Tellis (2018), point out the overall poor forecasting ability, the unstable parameter estimates and the difficulty to define a clear stopping rule for the time window regarding data collection for the Bass model (since the data should, in theory, end when the entire market has adopted). We add one more critique to the list: the parameter $m$, in the logic of the Bass model, appears to be in the DNA of the process since the onset and to be immutable over time. In all versions of the model that we have considered, the "shocks" (i.e., the unexpected citing papers) have no permanent effect in the sense that they determine, at most, a persistent (but transitory) departure from a path, which eventually leads to $m$. In the spirit of the unit roots literature, so much inspired by the contribution of Katarina Juselius and Søren Johansen, our suggestion for further methodological research is to try and conceive a new model, where the shocks are allowed to have a permanent effect on the "ceiling". We think that this is absolutely needed in the applications, such as the bibliometric ones, where the notion of "population at risk" or "potential" is not obvious. However, also in marketing, or epidemiology, or in the analysis of technological innovation, the final diffusion is likely to be influenced in a crucial way by events that are largely unpredictable; therefore, pretending that the same differential equation-where $m$ is fixed since $t=0$-drives the dynamics of the process along its entire history might not be a realistic representation of the observed phenomena.

Author Contributions: The authors contributed equally to the paper. All authors have read and agreed to the published version of the manuscript.
Funding: This research received no external funding.
Data Availability Statement: The data are available from the authors upon request.
Conflicts of Interest: The authors declare no conflict of interest.

## Appendix A. Bass vs. Autoregressive Models

It is interesting to observe that (12) can be regarded as a univariate $\mathrm{AR}(1)$ model for $C_{t}$ with state dependent parameters. In fact, the model can be rewritten as the following:

$$
\begin{equation*}
\Delta C_{t}=\mu+\pi_{t} C_{t-1}+u_{t} \tag{A1}
\end{equation*}
$$

with

$$
\begin{aligned}
\mu & =p m \\
\pi_{t} & =-p+q\left(1-\frac{C_{t-1}}{m}\right)
\end{aligned}
$$

$m$ can be seen as a steady state for $C_{t}$ : in fact, if $C_{t-1}=m$, then $\pi_{t}=-p$, so that in the absence of shocks (i.e., $u_{t}=0$ ), we have that $\Delta C_{t}=p m-p m=0$. The state dependent parameter $\pi_{t}$ controls the strength of the adjustment to the steady state.

- If $q=0, p>0$, and $m>0$, then $\pi_{t}=-p$ and the model collapses into a standard stationary $\operatorname{AR}(1)$ with unconditional expectation $m$.
- If $q>p>0$, and $m>0$, then $\pi(0)>0$ so that initially the system behaves like an explosive $\operatorname{AR}(1)$ with a positive drift $m p$. When $C_{t-1}=m\left(1-\frac{p}{q}\right)$, then $\pi_{t}=0$ so that the system locally behaves like a random walk with drift. When $C_{t-1}>m\left(1-\frac{p}{q}\right)$, then $\pi_{t}<0$ so that the system starts adjusting. An illustrative example, based on the estimated parameters for the methodological index, is given in Figure A1.


Figure A1. Illustration of $\pi_{t}$ as a function of $C_{t-1}$, with $p=0.00549, q=0.0242, m=1227.7$ as in the estimated equation for $C_{M, t}$.

Similarly, the model (19) can be seen as a univariate AR(2) model for $C_{t}$ with state dependent parameters:

$$
\begin{equation*}
\Delta C_{t}=\mu+\pi_{t} C_{t-1}+\gamma \Delta C_{t-1}+u_{t} \tag{A2}
\end{equation*}
$$

with:

$$
\begin{aligned}
\mu & =-\alpha m p \\
\pi_{t} & =\alpha\left(p-q\left(1-\frac{C_{t-1}}{m}\right)\right) \\
\gamma & =1+\alpha
\end{aligned}
$$

so that when $\alpha=-1$ (A2) collapses into (A1). We remark that as far as $\alpha$ is negative the sign of $\pi_{t}$ is the same in both models, and depends only on the sign of $C_{t-1}-m\left(1-\frac{p}{q}\right)$. The magnitude of $\pi_{t}$ instead is affected by $\alpha$ : everything else being fixed, when $-1<\alpha<0$, the process is less explosive at the beginning, and the strength of adjustment is weaker in the end, as compared to the case $\alpha=-1$.

Finally, the model (28) can be seen as a $\operatorname{VAR}(2)$ model for $C_{t}=\left[C_{1, t}, \ldots, C_{n, t}\right]^{\prime}$ with state-dependent parameters:

$$
\begin{equation*}
\Delta C_{t}=\mu+\alpha \boldsymbol{\beta}_{t}^{\prime} \boldsymbol{C}_{t-1}+\Gamma \Delta \boldsymbol{C}_{t-1}+\boldsymbol{u}_{t} \tag{A3}
\end{equation*}
$$

where

$$
\begin{aligned}
\underset{n \times 1}{\boldsymbol{C}_{t}} & =\left[\begin{array}{c}
C_{1, t} \\
\vdots \\
C_{n, t}
\end{array}\right], \underset{n \times 1}{\boldsymbol{u}_{t}}=\left[\begin{array}{c}
u_{1, t} \\
\vdots \\
u_{n, t}
\end{array}\right], \underset{n \times 1}{\boldsymbol{\mu}}=-\left[\begin{array}{c}
\sum_{j=1}^{n} \alpha_{1 j} p_{j} m_{j} \\
\vdots \\
\sum_{j=1}^{n} \alpha_{n j} p_{j} m_{j}
\end{array}\right] \\
\underset{n \times n}{\boldsymbol{\alpha}} & =\left[\begin{array}{ccc}
\alpha_{11} & \cdots & \alpha_{1 n} \\
\vdots & \ddots & \vdots \\
\alpha_{n 1} & \cdots & \alpha_{n n}
\end{array}\right], \underset{n \times n}{\boldsymbol{\beta}_{t}}=\operatorname{diag}\left\{p_{i}-q_{i}\left(1-\frac{C_{i, t-1}}{m_{i}}\right)\right\}, \quad \boldsymbol{\Gamma}=\boldsymbol{I}_{n}+\boldsymbol{\alpha} .
\end{aligned}
$$

It is easily seen that, when $\alpha$ is diagonal, (A3) collapses into $n$ seemingly unrelated equations such as (A2), while when $\alpha=-\boldsymbol{I}_{n}$, (A3) collapses into $n$ seemingly unrelated equations such as (A1).

## Appendix B. Sensitivity to $\omega$

As discussed in Section 2, the parameter $\omega$ in (1) and (2) controls for the weight of the "Mainly Applied" (MA) and "Mainly Methodological" (MM) papers on the aggregate indices $c_{1, t}$ (applied) and $c_{2, t}$ (methodological). Meaningful values of $\omega$ are in the range $(0.5 ; 1)$ : with $\omega=0.5$ the papers classified as MA and MM are essentially pooled together, and allowed to contribute evenly to both indices. In the opposite polar case, $\omega=1, \mathrm{MA}$ (or MM ) is considered equivalent to PA (or PM). We observe that, in principle, instead of a single weight $\omega$, it would be possible to consider two different weights for MA and MM papers, for example, $\omega_{A}$ and $\omega_{M}$, defining the following:

$$
\begin{aligned}
& c_{1, t}=c_{P A, t}+\omega_{A} c_{M A, t}+\left(1-\omega_{M}\right) c_{M M, t} \\
& c_{2, t}=c_{P M, t}+\omega_{M} c_{M M, t}+\left(1-\omega_{A}\right) c_{M A, t}
\end{aligned}
$$

We remark, however, that in our dataset, any value $0.5<\omega_{M}<1$ would leave the two indexes essentially unchanged since there are only 92 papers classified as MM in front of 716 classified as PM and 4198 classified as PA; therefore, our choice to set $\omega_{M}=\omega_{A}$ is a minor problem. Conversely, in our dataset, the critical issue is $\omega_{A}$, mainly because of its impact on $c_{2, t}$ : in fact, there are 1451 papers classified as MA and 716 classified as PM, so that setting $\omega_{A}=0.5$, the MA papers would be as influential as the PM papers in the index $c_{2, t}$. This argument induced us to set $\omega_{A}=\omega_{M}=0.85$. With this choice, $\left(1-\omega_{A}\right)$ is relatively close to 0 so that $c_{2, t}$ reflects mainly the 716 PM papers and, therefore, is a more reliable measure of the methodological research. Notice that this choice has a minor impact on the reliability of the applied index $c_{1, t}$ for two reasons: (i) the 4198 papers classified as PA outnumber the 1451 MA papers, and (ii) the correlation between $c_{P A, t}$ and $c_{M A, t}$ it quite high, $69.5 \%$ (see Figure 1) (conversely the correlation between $c_{P M, t}$ and $c_{M A, t}$ is only $1.7 \%$ ). Finally, notice that setting $\omega_{A}=0.85$ would make it approximately equal to $n_{P A} /\left(n_{P A}+n_{P M}\right)=0.8545$ : in practice, this corresponds to the assumption that the share of "applied research" of an MA paper is similar, on average, to the share of applied research in the econometric literature referring to KJ and SJ papers in general. Let us now discuss how a different choice of $\omega$ would affect our results.

As illustrated in Table A1, changing $\omega$ affects quite relevantly the magnitude of the indices (especially $c_{2, t}$ ), as well as the correlation among them. To explain the impact on the magnitude, remember that, when $\omega=1$, the 1451 MA papers are treated de facto as the "Purely Applied"(PA) ones, whereas when $\omega=0.5$, only half of them (725.5) is treated as applied, while the other half is treated as methodological, and therefore, contribute also to the methodological index $c_{2, t} \cdot 25$ The impact on the correlation is instead explained by the fact that the correlation between $c_{P A, t}$ and $c_{M A, t}$ it quite high ( $69.5 \%$ ), whereas the correlation between $c_{P M, t}$ and $c_{M A, t}$ is negligible (1.7\%); see Figure 1.

Table A1. Sensitivity to $\omega$ : impact of $\omega$ on some characteristics of the composite citation indices.

|  | $\boldsymbol{\omega}=\mathbf{0 . 5}$ | $\boldsymbol{\omega}=\mathbf{0 . 8 5}$ | $\boldsymbol{\omega}=\mathbf{1}$ |
| :---: | :---: | :---: | :---: |
| $C_{1, T}$ | 4969.5 | 5445.2 | 5649 |
| $C_{2, T}$ | 1487.5 | 1011.8 | 808 |
| $\operatorname{corr}\left(c_{1, t}, c_{2, t}\right)$ | 0.543 | 0.253 | 0.048 |

Given this impact of $\omega$ on the composite indices, it is interesting to analyze to which extent the results of the econometric model depend on it. The analysis is limited to the general model (29) since our analysis shows that it is preferable with respect to the restricted
counterparts (27) and (18). In this appendix, we show that our results are essentially robust to changes in $\omega$.

Table A2 shows how the estimates of the reduced for changes when $\omega$ is changed.
Table A2. Sensitivity to $\omega$ : ML estimates of the reduced form parameters based on model (29).

|  |  | $\omega=\mathbf{0 . 5}$ |  | $\omega=\mathbf{0 . 8 5}$ |  | $\omega=\mathbf{1}$ |  |
| :--- | :--- | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Estimate | t-Ratio | Estimate | t-Ratio | Estimate | t-Ratio |
| APP | $\hat{\alpha}_{1,1}$ | -0.620 | -7.17 | -0.508 | -6.26 | -0.459 | -5.92 |
|  | $\hat{\alpha}_{1,2}$ | 0.782 | 4.00 | 0.715 | 2.81 | 0.624 | 2.33 |
|  | $\hat{\beta}_{0,1}$ | 16.95 | 4.58 | 19.81 | 4.78 | 21.11 | 4.88 |
|  | $\hat{\beta}_{1,1}$ | 0.0209 | 5.13 | 0.0190 | 4.61 | 0.0180 | 4.33 |
|  | $\hat{\beta}_{2,1}$ | $-1.86 \times 10^{-6}$ | -2.06 | $-1.32 \times 10^{-6}$ | -1.60 | $-1.06 \times 10^{-6}$ | -1.33 |
|  | $\hat{\sigma}_{1}$ | 7.46 |  | 8.15 |  | 8.62 |  |
|  | $\hat{\alpha}_{2,1}$ | 0.0902 | 2.25 | 0.0635 | 2.34 | 0.0590 | 2.38 |
|  | $\hat{\alpha}_{2,2}$ | -0.545 | -6.12 | -0.547 | -6.44 | -0.596 | -7.04 |
|  | $\hat{\beta}_{0,2}$ | 8.87 | 4.81 | 6.74 | 4.65 | 5.64 | 4.42 |
|  | $\hat{\beta}_{1,2}$ | 0.0161 | 2.66 | 0.0187 | 2.73 | 0.0223 | 2.96 |
|  | $\hat{\beta}_{2,2}$ | $-9.54 \times 10^{-6}$ | -2.23 | $-1.97 \times 10^{-5}$ | -2.86 | $-3.15 \times 10^{-5}$ | -3.39 |
|  | $\hat{\sigma}_{2}$ | 3.46 |  | 2.72 |  | 2.71 |  |

The sign and significance of the parameters are essentially the same irrespective of $\omega$. Interestingly, the difference in the correlation between the two indices induced by $\omega$, illustrated in Table A1, are reflected in different estimates of $\rho$, whereas the estimates of $\alpha_{1,2}$ and $\alpha_{2,1}$ (i.e., the parameters controlling for the dynamic interaction among the two processes) remain quite stable. Due to this, the (unreported) pattern of the standardized IRs and cumulative IRs computed with $\omega=0.5$ and $\omega=1$ are very similar to those illustrated in Figures 4 and 5 for $\omega=0.85$. Unreported results show that also the misspecification tests are qualitatively unchanged for all $\omega$ s with respect to those reported in Table 5 for model (29): homoskedasticity and uncorrelatedness appear acceptable for any value of $\omega$.

Table A3 illustrate how the structural parameters change when $\omega$ is changed. The influence on the $m$ has an obvious interpretation: as $\omega$ increases, a larger share of the MA papers is removed from the methodological index (so that $m_{2}$ declines) and added to the applied index (so that $m_{1}$ increases). As for the $p$ s and the $q s$, we observe that as $\omega$ increases, $p_{1}$ and $q_{1}$ decrease, whereas $p_{2}$ and $q_{2}$ increase. As a consequence of these changes, the timing of the peaks, obtained by formula (9), change: specifically, as $\omega$ increases, the peak in the applied literature moves to the right, whereas the peak in the methodological peak moves to the left. The distance between the peaks is 13 years with $\omega=0.5$, and about 24 years when $\omega=1$. This is not surprising: as illustrated in Figure 1, the dynamic behavior of the MA paper resembles closely the PA papers, and therefore, when $50 \%$ of them are considered methodological, $c_{1, t}$ and $c_{2, t}$ become more similar, and the two peaks become closer (although they still remain quite far away from each other).

An interesting consequence of the fact that, increasing $\omega$, the peak of the applied literature moves ahead is that the quality of the structural parameters for the applied curve (already quite poor with $\omega=0.5$ ) decreases considerably: when $\omega=1$, the standard error associated to $\hat{m}_{1}$ is as large as 9516, and the standard error associated to the estimated timing of the peak $\hat{t}_{1}^{P}$ turns out to be 42 quarters, more than 10 years. It is a well-known fact in the literature that the estimates of the Bass model are quite poor if the sample period does not include the inflection point, which is quite likely the case for the applied literature if we trust the point estimates, and even more so when $\omega=1$.

Table A3. Sensitivity to $\omega$ : ML estimates of the structural form parameters based on model (29). The standard errors of the $\hat{t}^{P}$ measured in quarters.

|  |  | $\omega=\mathbf{0 . 5}$ |  | $\omega=\mathbf{0 . 8 5}$ |  | $\omega=\mathbf{1}$ |  |
| :--- | :--- | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Coefficient | Estimate | Std.err. | Estimate | Std.err. | Estimate | Std.err. |
|  | $\hat{h}_{1}$ | $11,994.6$ | 3644.9 | $15,350.6$ | 6396.2 | $18,058.0$ | 9516.0 |
|  | $\hat{p}_{1}$ | 0.00141 | $3.94 \times 10^{-4}$ | 0.00129 | $4.65 \times 10^{-4}$ | 0.00117 | $5.31 \times 10^{-4}$ |
|  | $\hat{q}_{1}$ | 0.0224 | 0.00425 | 0.0203 | 0.00442 | 0.0192 | 0.00456 |
|  | $t_{1}^{P}$ | $2018: 1$ | 22.6 | $2020: 4$ | 32.4 | $2023: 2$ | 42.4 |
|  | $c_{1}^{P}$ | 76 | 10.3 | 88 | 18.4 | 98 | 28.0 |
|  | $C_{1}^{P}$ | 5618 | 1719.8 | 7188 | 3050.6 | 8479 | 4578.2 |
|  |  |  |  |  |  |  |  |  |
|  | $m_{2}$ | 2129.1 | 351.5 | 1227.7 | 119.7 | 907.1 | 62.3 |
|  | $p_{2}$ | 0.00416 | $8.77 \times 10^{-4}$ | 0.00549 | $4.91 \times 10^{-4}$ | 0.00622 | 0.00133 |
| MET | $q_{2}$ | 0.0203 | 0.00604 | 0.0242 | 0.00643 | 0.0285 | 0.00682 |
|  | $t_{2}^{P}$ | $2005: 1$ | 11.7 | $2001: 2$ | 8.4 | $1999: 4$ | 7.2 |
|  | $c_{2}^{p}$ | 16 | 1.2 | 11 | 1.0 | 10 | 0.8 |
|  | $C_{2}^{p}$ | 846 | 122.6 | 475 | 49.9 | 355 | 34.2 |

## Notes

1 Many thanks for the provision of the initial Web of Science data to Evi Sachini, Antonis Kardasis and Penny Nikolaidou of the National Documentation Centre/N.H.R.F. based in Athens, Greece.
2 Around the same time Google Scholar (GS) reported more than 50,000 citations for the same 10 papers. We opted for WoS instead of GS because, to avoid double counting, the analysis carried on in this paper is based on the citing papers instead of the citations, and working out the citing papers from GS is not easy. Admittedly, one drawback with using WoS instead of GS is that books cannot be considered; we think however that this would not substantially change the picture. In fact, according to GS, the book Johansen (1995) would rank 4th in terms of citations, the book Juselius (2006) would rank 6th, and adding both books the total citations count would be about $15 \%$ higher; however, since many papers citing one of the books will cite also some of the older papers, the impact of the books on the citing papers is likely to be way less than $10 \%$.
3 Among the "super-citing" papers, there are also 14 papers with six citations and 53 papers with five citations. We remark that 40 of the 6457 papers (i.e., $0.62 \%$ ) are authored or coauthored by SJ and/or KJ: given the small share we did not correct for self-citations.
4 An alternative way of measuring the influence of a paper could be based on counting the authors instead of the papers. We could then consider the number of authors citing $K J$ or $S J$ in each quarter, or preferably the number of "new authors", i.e., the number of authors citing KJ or SJ for the first time in each quarter, who never cited them before (this would avoid double counting, and would be a more precise measure of "contagion"). We do not explore this alternative in the present paper, leaving it for future research.
5 Classifying an econometric paper as "methodological" or "applied" is clearly arbitrary to some extent. A general discussion, although related to the 'delineation of scientific areas' may be found in Zitt (2006); he states that fields may be defined at various levels (e.g., institutional setting of academic actors; shared topics and possibly shared journals; shared terminology; close connections of collaboration or citation, etc.) and concludes that ". . . natural borders, generally speaking, are an illusion" (Zitt 2006, p. 6). In fact, a more scientific-bibliometric related methodological approach could be the analysis based on networks, as for instance in Vieira and Teixeira (2010), although this is outside the scope of the present paper.
6 When unsure regarding the screening, we proceeded following Katsaliaki and Mustafee (2011, p. 1434): "The two authors independently and critically reviewed all the abstracts of the (...) papers and read the full text when necessary." Notice that an alternative classification scheme could be based on the publishing journal since some journals are more oriented toward applications, while others are more methodological. As discussed below, we believe that our approach provides a more accurate measure.
7 The title of the MA paper by Baillie and Bollerslev is "Common stochastic trends in a system of exchange rates", while the title of the PM paper by Gilbert is "Economic theory and econometric models"
8 Actually, at the individual level, the term "innovator" associated to a constant hazard is somewhat misleading, and not exactly a synonym of "early adopter". In fact, an individual with constant hazard rate might well be a laggard, especially if his/her individual hazard rate is low. The parameter $p$ is hardly interpretable in epidemiology, where the notion of "innovator" is essentially limited to the "patient zero".
$9 \quad$ We remark that that the solution is not unique. The formulae in (15) are the ones giving positive values of $m, p$ and $q$ with our estimated $\beta^{\prime}$ 's.

Maintaining the assumption that $u_{t}$ is i.i.d. normal, an alternative estimation strategy could be based on Non Linear Least Squares (NLLS). Estimates of $m, p$ and $q$ would be based on the following:

$$
\min _{m, p, q} \sum_{t=2}^{T}\left(c_{t}-m p-(q-p) C_{t-1}+\frac{q}{m} C_{t-1}^{2}\right)^{2}
$$

The advantage of NLLS is that it provides directly the estimates of the parameters of interest ( $m, p$ and $q$ ) and the corresponding standard error, without having to resort to the delta method. The disadvantage is that convergence of the numerical optimization routines is sometimes not easy: this is partly due to the strong collinearity, and partly to the fact that the optimization problem has two solutions. In the following, we opt for OLS and the delta method.
We decided to adopt slightly different symbols with respect to BF. In particular our $\alpha$ has opposite sign with respect to theirs. The parametrization (23) is also better suited than (24) for multivariate generalizations, as illustrated in the next subsection. Actually, Boswijk et al. (2009) propose an heteroskedastic version of the model, where $u_{t}=\operatorname{diag}\left\{c_{i, t-1}^{\phi}\right\} \varepsilon_{t}$, with $\varepsilon_{t}=$ $\left[\varepsilon_{1, t}, \ldots, \varepsilon_{n, t}\right] \sim \operatorname{iid} N_{n}\left(0, \Omega_{\varepsilon}\right)$ and $\phi$ fixed to either $1 / 2$ or 1 . In this paper we only briefly discuss the heteroskedastic BFF model, since in our application suitable heteroskedasticity tests seem to accept the hypothesis of homoskedasticity. Precisely,

$$
\underset{\left(3 n^{2}+n\right) \times 3 n}{\boldsymbol{H}_{\beta}}=\operatorname{diag}\left\{\boldsymbol{H}_{i}\right\}, \quad \underset{\left(3 n^{2}+n\right) \times 1}{\boldsymbol{h}_{\beta}}=\operatorname{diag}\left\{\boldsymbol{h}_{i}\right\} \mathbf{1}_{n},
$$

with

$$
\underset{(3 n+1) \times 3}{\boldsymbol{H}_{i}}=\left[\begin{array}{cc}
\boldsymbol{u}_{n, i} \otimes\left[\boldsymbol{u}_{3,2}, \boldsymbol{u}_{3,3}\right] & \mathbf{0}_{3 n, 1} \\
\mathbf{0}_{1,2} & 1
\end{array}\right], \underset{(3 n+1) \times 1}{\boldsymbol{h}_{i}}=\left[\begin{array}{c}
\boldsymbol{u}_{n, i} \otimes \boldsymbol{u}_{3,1} \\
0
\end{array}\right] \quad i=1, \ldots, n .
$$

A diagonal $\alpha$ corresponds to $\boldsymbol{H}_{\boldsymbol{\alpha}}=\operatorname{diag}\left\{\boldsymbol{u}_{n, i}\right\}$. As already observed, in this case ML would not correspond to equation by equation OLS, due to the correlation of the error terms. One might maximize the likelihood either by iterated SUR as illustrated in Section 3.2, or equivalently using the algorithm illustrated here.
Minor modifications are needed if instead we assume heteroskedasticity of the type postulated in Boswijk et al. (2009), where $\boldsymbol{u}_{t}=\boldsymbol{W}_{t} \varepsilon_{t}$, with $\boldsymbol{W}_{t}=\operatorname{diag}\left\{c_{i, t-1}^{\phi}\right\}$ ( $\phi$ is assumed to be known) and $\varepsilon_{t} \sim \operatorname{iid} N\left(\mathbf{0}, \boldsymbol{\Omega}_{\varepsilon}\right)$. Notice that, premultiplying (30), left and right, by $W_{t}^{-1}$, using the properties of the vec operator, one obtains either the following:

$$
\boldsymbol{W}_{t}^{-1} \boldsymbol{Y}_{t}=\left[\left(\boldsymbol{X}_{t-1}^{\prime} \otimes \boldsymbol{W}_{t}^{-1}\right)\left(\boldsymbol{\beta} \otimes \boldsymbol{I}_{n}\right)\right] \operatorname{vec}(\boldsymbol{\alpha})+\boldsymbol{u}_{t}
$$

or the following:

$$
\boldsymbol{W}_{t}^{-1} \boldsymbol{Y}_{t}=\left[\left(\boldsymbol{X}_{t-1}^{\prime} \otimes \boldsymbol{W}_{t}^{-1}\right)\left(\boldsymbol{I}_{3 n+1} \otimes \boldsymbol{\alpha}\right)\right] \operatorname{vec}\left(\boldsymbol{\beta}^{\prime}\right)+\boldsymbol{u}_{t}
$$

The first equation allows to estimate $\alpha$ by GLS when $\beta$ and $\Omega$ are known, while the second allows to estimate $\beta$ by GLS when $\alpha$ and $\Omega$ are known. A "switching" iterative algorithm similar to Hansen (2003) is therefore possible also in this case. Of course, linear restrictions on $\operatorname{vec}(\boldsymbol{\alpha})$ or $\operatorname{vec}(\boldsymbol{\beta})$ are easily dealt with also in this case.
Models (18) and (27) are also estimated using iterated SURE, obtaining exactly the same results.
We also considered different values of $k$, from 4 to 20, and the results remain essentially unchanged. Regarding the number of degrees of freedom, as illustrated in Appendix A, the standard Bass model can be seen as an AR(1) with state dependent parameters, while the BF and BFF models can be seen as $\operatorname{AR}(2)$ : therefore we considered heuristically $k-p$ degrees of freedom in the $Q$ test, with $p=1$ for the standard Bass model and $p=2$ for BF and BFF models.
Actually, the slope in the auxiliary regression is negative in some cases, which is exactly the opposite of BF intuition. We think that the result might reflect the neglected autocorrelation rather than heteroskedasticity: the ample swings in the residuals clearly visible in Figure 3 are misinterpreted by the test as heteroskedasticity.
Alternative initializations are possible: this point is further discussed in footnote 25 .
For simplicity, we do not "orthogonalize" the shocks by assuming some direction for the simultaneous relationship: we believe that this is justified in this case, given the modest correlation between the residuals ( $16.1 \%$ ). As a robustness check we also tried to orthogonalize in either direction, and to apply the "ordering invariant" method proposed in Pesaran and Shin (1998) but, as expected given the low correlation, the results are essentially unchanged. For a discussion of the simultaneous correlation, see also Appendix B.
It is important to remark that, given the nonlinear dynamics implied by (29), the impulse responses will change according to the initial conditions. We also considered alternative initializations, starting in different points of the diffusion path: we observed that when the impulse is given further ahead along the diffusion path, the shape of the responses changes in a rather
intuitive way: the peak of the cumulative IRs occurs earlier, and the intensity becomes weaker. This can be explained in the light of the discussion presented in Appendix A: in the initial stages of the process, when both $C_{1, t}$ and $C_{2, t}$ are close to zero and much lower than $m_{1}$ and $m_{2}$, respectively, the processes behave as explosive $\operatorname{AR}(2)$, and therefore, the shocks are initially amplified; however, as $C_{1, t}$ and $C_{2, t}$ grow, the processes become less and less explosive, until eventually they start adjusting and the cumulative impact of the shock is driven down to zero. However, some characteristics of the cumulative IRs do not change, even when the initial conditions are modified: the cumulative cross impact seems to be relatively stronger from the methodological to the applied literature than vice versa.
23 Defining $\psi=\left[t^{P}, \bar{c}^{P}, \bar{C}^{P}\right]^{\prime}$, starting from (9)-(11), we have the following:

$$
J_{\psi \cdot \theta}=\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}^{\prime}}=\left[\begin{array}{ccc}
(p+q)^{2} & 0 & 0 \\
0 & 4 q & 0 \\
0 & 0 & 2 q
\end{array}\right]^{-1}\left[\begin{array}{ccc}
0 & \ln p-\ln q-1-\frac{q}{p} & \ln p-\ln q+1+\frac{p}{q} \\
(p+q)^{2} & 2 m(p+q) & \frac{m}{q}(p+q)(q-p) \\
q-p & -m & m \frac{p}{q}
\end{array}\right]
$$

The variance-covariance matrix for $\hat{\psi}$ is then obtained as the following:

$$
\hat{\Sigma}_{\hat{\psi}}=\hat{J}_{\psi \cdot \theta} \hat{\Sigma}_{\hat{\theta}} \hat{\jmath}_{\psi \cdot \theta}^{\prime} .
$$

These estimates are based on model (27) since we could not achieve convergence in model (29). The estimates based on model (18) and the same sample are almost identical.

Similarly, when $\omega=1$, the 92 MM papers are entirely treated as methodological, whereas, when $\omega=0.5$, only half of them (46) are treated as methodological, while the other half is treated as applied. Given the small number of MM papers, their influence on the indices is negligible, and that is why in our discussion we emphasize the role of the MA papers.

## References

Bass, Frank M. 1969. A new product growth for model consumer durables. Management Science 15: 215-27. [CrossRef]
Bjork, Samuel, Avner Offer, and Gabriel Söderberg. 2014. Time series citation data: The nobel prize in economics. Scientometrics 98: 185-96. [CrossRef]
Boswijk, H. Peter, Dennis Fok, and Philip Hans Franses. 2009. A New Multivariate Product Growth Model. Technical Report, Econometrics Discussion Paper 2009/07. Amsterdam: University of Amsterdam.
Boswijk, H. Peter, and Philip Hans Franses. 2005. On the econometrics of the bass diffusion model. Journal of Business \& Economic Statistics 23: 255-68. [CrossRef]
Boswijk, H. Peter, Philip Hans Franses, and Dick van Dijk. 2010. Cointegration in a historical perspective. Journal of Econometrics 158: 156-59. [CrossRef]
Chandrasekaran, Deepa, and Gerard J. Tellis. 2018. A Summary and Review of New Product Diffusion Models and Key Findings. Cheltenham: Edward Elgar Publishing, chapter 14, pp. 291-312. [CrossRef]
Coupé, Tom. 2003. Revealed performances: Worldwide rankings of economists and economics departments, 1990-2000. Journal of the European Economic Association 1: 1309-45. [CrossRef]
Engle, Robert F., and Clive W. J. Granger. 1987. Co-integration and error correction: Representation, estimation, and testing. Econometrica 55: 251-71. [CrossRef]
Eryarsoy, Enes, Dursun Delen, Behrooz Davazdahemami, and Kazim Topuz. 2021. A novel diffusion-based model for estimating cases, and fatalities in epidemics: The case of covid-19. Journal of Business Research 124: 163-87. [CrossRef]
Fok, Dennis, and Philip Hans Franses. 2007. Modeling the diffusion of scientific publications. Journal of Econometrics 139: 376-90. [CrossRef]
Franses, Philip Hans. 2003. The diffusion of scientific publications: The case of econometrica, 1987. Scientometrics 56: 29-42. [CrossRef]
Garfield, Eugene, Morton V. Malin, and Henry Small. 1978. Citation data as science indicators. In Toward a Metric of Science: The Advent of Science Indicators. Edited by Yehuda Elkana, Joshua Lederberg, Robert K. Merton, Arnold Thackray and Harriet Zuckerman. New York: John Wiley \& Sons, pp. 179-207.
Guseo, Renato, and Mariangela Guidolin. 2008. Cellular automata and riccati equation models for diffusion of innovations. Statistical Methods and Applications 17: 291-308. [CrossRef]
Hansen, Peter Reinhard. 2003. Structural changes in the cointegrated vector autoregressive model. Journal of Econometrics 114: 261-95. [CrossRef]
Heeler, Roger M., and Thomas P. Hustad. 1980. Problems in predicting new product growth for consumer durables. Management Science 26: 1007-20. [CrossRef]
Hendry, David, and Katarina Juselius. 2001. Explaining cointegration analysis. part ii. The Energy Journal 22: 75-120. [CrossRef]
Hyman, Michael R. 1988. The timeliness problem in the application of bass-type new product-growth models to durable sales forecasting. Journal of Business Research 16: 31-47. [CrossRef]
Johansen, Søren. 1988. Statistical analysis of cointegration vectors. Journal of Economic Dynamics and Control 12: 231-54. [CrossRef]
Johansen, Søren. 1991. Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. Econometrica 59: 1551-80. [CrossRef]

Johansen, Søren. 1992a. Cointegration in partial systems and the efficiency of single-equation analysis. Journal of Econometrics 52: 389-402. [CrossRef]
Johansen, Søren. 1992b. Determination of cointegration rank in the presence of a linear trend. Oxford Bulletin of Economics and Statistics 54: 383-97. [CrossRef]
Johansen, Søren. 1992c. Testing weak exogeneity and the order of cointegration in uk money demand data. Journal of Policy Modeling 14: 313-34. [CrossRef]
Johansen, Søren. 1995. Likelihood-Based Inference in Cointegrated Vector Autoregressive Models. Oxford: Oxford University Press. [CrossRef]
Johansen, Søren. 2006. Statistical analysis of hypotheses on the cointegrating relations in the i(2) model. Journal of Econometrics 132: 81-115. [CrossRef]
Johansen, Søren, and Katarina Juselius. 1990. Maximum likelihood estimation and inference on cointegration—With applications to the demand for money. Oxford Bulletin of Economics and Statistics 52: 169-210. [CrossRef]
Johansen, Søren, and Katarina Juselius. 1992. Testing structural hypotheses in a multivariate cointegration analysis of the ppp and the uip for uk. Journal of Econometrics 53: 211-44. [CrossRef]
Johansen, Søren, and Katarina Juselius. 1994. Identification of the long-run and the short-run structure an application to the islm model. Journal of Econometrics 63: 7-36. [CrossRef]
Johansen, Søren, Rocco Mosconi, and Bent Nielsen. 2000. Cointegration analysis in the presence of structural breaks in the deterministic trend. The Econometrics Journal 3: 216-49. [CrossRef]
Juselius, Katarina. 2006. The Cointegrated VAR Model: Methodology and Applications. Oxford: Oxford University Press.
Juselius, Katarina. 2021. Searching for a theory that fits the data: A personal research odyssey. Econometrics 9: 5. [CrossRef]
Kalaitzidakis, Pantelis, Theofanis P. Mamuneas, and Thanasis Stengos. 1999. European economics: An analysis based on publications in the core journals. European Economic Review 43: 1150-68. [CrossRef]
Kalaitzidakis, Pantelis, Theofanis P. Mamuneas, and Thanasis Stengos. 2003. Rankings of academic journals and institutions in economics. Journal of the European Economic Association 1: 1346-66. [CrossRef]
Katsaliaki, Korina, and Navonil Mustafee. 2011. Applications of simulation within the healthcare context. Journal of the Operational Research Society 62: 1431-51. [CrossRef]
Koop, Gary, M. Hashem Pesaran, and Simon M. Potter. 1996. Impulse response analysis in nonlinear multivariate models. Journal of Econometrics 74: 119-47. [CrossRef]
Kuc-Czarnecka, Marta, Samuele Lo Piano, and Saltelli Andrea. 2020. Quantitative storytelling in the making of a composite indicator. Social Indicators Research 1: 775-802. [CrossRef]
Lütkepohl, Helmut. 2016. Impulse Response Function. London: Palgrave Macmillan UK, pp. 1-5. [CrossRef]
Min, Chao, Ying Ding, Jiang Li, Yi Bu, Lei Pei, and Jianjun Sun. 2018. Innovation or imitation: The diffusion of citations. Journal of the Association for Information Science and Technology 69: 1271-82. [CrossRef]
Nardo, Michela, Michaela Saisana, Andrea Saltelli, Stefano Tarantola, Anders Hoffmann, and Enrico Giovannini. 2008. Handbook on Constructing Composite Indicators: Methodology and User Guide. Paris: OECD Publishing.
Pesaran, H. Hashem, and Yongcheol Shin. 1998. Generalized impulse response analysis in linear multivariate models. Economics Letters 58: 17-29. [CrossRef]
Potter, Simon M. 2000. Nonlinear impulse response functions. Journal of Economic Dynamics and Control 24: 1425-46. [CrossRef]
Redner, Sidney. 1998. How popular is your paper? An empirical study of the citation distribution. The European Physical Journal B-Condensed Matter and Complex Systems 4: 131-34. [CrossRef]
Satoh, Daisuke. 2001. A discrete bass model and its parameter estimation. Journal of the Operations Research Society of Japan 44: 1-18. [CrossRef]
Srinivasan, V., and Charlotte H. Mason. 1986. Nonlinear least squares estimation of new product diffusion models. Marketing Science 5: 169-78. [CrossRef]
Stigler, Stephen M. 1994. Citation patterns in the journals of statistics and probability. Statistical Science 9: 94-108. [CrossRef]
Stokes, Donald E. 1997. Pasteur's Quadrant: Basic Science and Technological Innovation. Washington: Brookings Institution Press.
Van den Bulte, Christophe, and Gary L. Lilien. 1997. Bias and systematic change in the parameter estimates of macro-level diffusion models. Marketing Science 16: 338-53. [CrossRef]
Vieira, Pedro, and Aurora Teixeira. 2010. Are finance, management, and marketing autonomous fields of scientific research? An analysis based on journal citations. Scientometrics 85: 627-46. [CrossRef]
Zitt, Michel. 2006. Scientometric indicators: A few challenges. data mine-clearing; knowledge flows measurements; diversity issues. Invited plenary talk. Paper presented at the International Workshop on Webometrics, Informetrics and Scientometrics \& Seventh COLLNET Meeting, Nancy, France, December 12-15.

Article

# Cointegration and Adjustment in the CVAR( $\infty$ ) Representation of Some Partially Observed CVAR(1) Models 

Søren Johansen<br>Department of Economics, University of Copenhagen, Øster Farimagsgade 5, building 26, DK-1353 Copenhagen K, Denmark; Soren.Johansen@econ.ku.dk; Tel.: +45-35323071

Received: 18 September 2018; Accepted: 8 January 2019; Published: 10 January 2019


#### Abstract

A multivariate CVAR(1) model for some observed variables and some unobserved variables is analysed using its infinite order CVAR representation of the observations. Cointegration and adjustment coefficients in the infinite order CVAR are found as functions of the parameters in the CVAR(1) model. Conditions for weak exogeneity for the cointegrating vectors in the approximating finite order CVAR are derived. The results are illustrated by two simple examples of relevance for modelling causal graphs.


Keywords: adjustment coefficients; cointegrating coefficients; CVAR; causal models
JEL Classification: C32

## 1. Introduction

In a conceptual exploration of long-run causal order, Hoover (2018) applies the CVAR(1) model for the processes $X_{t}=\left(x_{1 t}, \ldots, x_{p t}\right)^{\prime}$ and $T_{t}=\left(T_{1 t}, \ldots, T_{m t}\right)^{\prime}$, to model a causal graph. The process $\left(X_{t}^{\prime} ; T_{t}^{\prime}\right)^{\prime}$ is a solution to the equations

$$
\begin{align*}
& \Delta X_{t+1}=M X_{t}+C T_{t}+\varepsilon_{t+1}  \tag{1}\\
& \Delta T_{t+1}=\eta_{t+1}
\end{align*}
$$

where the error terms $\varepsilon_{t}$ are independent identically distributed (i.i.d.) Gaussian variables with mean 0 and variance $\Omega_{\varepsilon}=\operatorname{diag}\left(\omega_{11}, \ldots, \omega_{p p}\right)>0$, and are independent of the errors $\eta_{t}$, which are (i.i.d.) Gaussian with mean 0 and variance $\Omega_{\eta}$.

Thus, the stochastic trends, $T_{t}$ are nonstationary random walks and conditions will be given below for $X_{t}$ to be $I(1)$, that is, nonstationary, but $\Delta X_{t}$ stationary. This will imply that $M X_{t}+C T_{t}$ is stationary, so that $X_{t}$ and $T_{t}$ cointegrate.

The entry $M_{i j} \neq 0$ means that $x_{j}$ causes $x_{i}$, which is written $x_{j} \rightarrow x_{i}$, and $C_{i j} \neq 0$ means that $T_{j} \rightarrow x_{i}$, and it is further assumed that $M_{i i} \neq 0$. Note that the model assumes that there are no causal links from $X_{t}$ to $T_{t}$, so that $T_{t}$ is strongly exogenous.

A simple example for three variables, $x_{1}, x_{2}, x_{3}$, and a trend $T$, is the graph

$$
T \rightarrow x_{1} \rightarrow x_{2} \rightarrow x_{3}
$$

where the matrices are given by

$$
M=\left(\begin{array}{ccc}
* & 0 & 0 \\
* & * & 0 \\
0 & * & *
\end{array}\right), C=\left(\begin{array}{l}
* \\
0 \\
0
\end{array}\right)
$$

where $*$ indicates a nonzero coefficient.
Provided that $I_{p}+M$ has all eigenvalues in the open unit disk, it is seen that

$$
M X_{t+1}+C T_{t+1}=\left(I_{p}+M\right)\left(M X_{t}+C T_{t}\right)+M \varepsilon_{t+1}+C \eta_{t+1}
$$

determines a stationary process defined for all $t$. We define a nonstationary solution to (1) for $t=0,1, \ldots$ by

$$
\begin{equation*}
X_{t}=-M^{-1} C \sum_{i=1}^{t} \eta_{i}+M^{-1} \sum_{i=0}^{\infty}\left(I_{p}+M\right)^{i}\left(M \varepsilon_{t-i}+C \eta_{t-i}\right) \text { and } T_{t}=\sum_{i=1}^{t} \eta_{i} . \tag{2}
\end{equation*}
$$

Note that the starting values are

$$
X_{0}=M^{-1} \sum_{i=0}^{\infty}\left(I_{p}+M\right)^{i}\left(M \varepsilon_{-i}+C \eta_{-i}\right) \text { and } T_{0}=0 .
$$

It is seen that $\Delta X_{t+1}, \Delta T_{t+1}$ and $M X_{t}+C T_{t}$ are stationary processes for all $t$, and that $\left(X_{t}^{\prime} ; T_{t}^{\prime}\right)^{\prime}$ is a solution to Equation (1). In the following, we assume that $\left(X_{t}^{\prime} ; T_{t}^{\prime}\right)^{\prime}$ is defined by (2) for $t=0,1, \ldots$

The paper by Hoover gives a detailed and general discussion of the problems of recovering causal structures from nonstationary observations $X_{t}$, or subsets of $X_{t}$, when $T_{t}$ is unobserved, that is, $X_{t}=\left(X_{1 t}^{\prime} ; X_{2 t}^{\prime}\right)^{\prime}$ where the observations $X_{1 t}$ are $p_{1}$-dimensional and the unobserved processes $X_{2 t}$ and $T_{t}$ are $p_{2}$ - and $m$-dimensional respectively, $p=p_{1}+p_{2}$. It is assumed that there are at least as many observations as trends, that is $p_{1} \geq m$.

Model (1) is therefore rewritten as

$$
\begin{align*}
& \Delta X_{1, t+1}=M_{11} X_{1 t}+M_{12} X_{2 t}+C_{1} T_{t}+\varepsilon_{1, t+1} \\
& \Delta X_{2, t+1}=M_{21} X_{1 t}+M_{22} X_{2 t}+C_{2} T_{t}+\varepsilon_{2, t+1}  \tag{3}\\
& \Delta T_{t+1}=\eta_{t+1}
\end{align*}
$$

Note that there is now a causal link from the observed process $X_{1 t}$ to the unobserved process $X_{2 t}$ if $M_{21} \neq 0$.

It follows from (3) that $X_{1 t}$ is $I(1)$ and cointegrated with $p_{1}-m$ cointegrating vectors $\beta$, see Theorem 1. Therefore, $\Delta X_{1 t}$ has an infinite order autoregressive representation, see (Johansen and Juselius 2014, Lemma 2), which is written as

$$
\begin{equation*}
\Delta X_{1, t+1}=\alpha \beta^{\prime} X_{1 t}+\sum_{i=1}^{\infty} \Gamma_{i} \Delta X_{1, t+1-i}+v_{t+1^{\prime}}^{\beta} \tag{4}
\end{equation*}
$$

where the operator norm $\left\|\Gamma_{i}\right\|=\lambda_{\max }^{1 / 2}\left(\Gamma_{i}^{\prime} \Gamma_{i}\right)$ is $O\left(\rho^{i}\right)$ for some $0<\rho<1$. The matrices $\alpha$ and $\beta$ are $p_{1} \times m$ of rank $m$, and $v_{t+1}^{\beta}=\Delta X_{1, t+1}-E\left(\Delta X_{1, t+1} \mid \mathcal{F}_{t}^{\beta}\right)$, where $\mathcal{F}_{t}^{\beta}=\sigma\left(\Delta X_{1 s}, s \leq t, \beta^{\prime} X_{1 t}\right)$. Thus, $X_{1 t}$ is not measurable with respect to $\mathcal{F}_{t}^{\beta}$, but $\beta^{\prime} X_{1 t}$ is measurable with respect to $\mathcal{F}_{t}^{\beta}$. Here, the prediction errors $v_{t+1}^{\beta}$ are i.i.d. $N_{p_{1}}(0, \Sigma)$, where $\Sigma$ is calculated below. The representation of $X_{1 t}$, similar to (2), is

$$
\begin{equation*}
X_{1 t}=\beta_{\perp}\left(\alpha_{\perp}^{\prime} \Gamma \beta_{\perp}\right)^{-1} \alpha_{\perp}^{\prime} \sum_{i=1}^{t} v_{i}^{\beta}+\sum_{i=0}^{\infty} C_{i} v_{t-i}^{\beta}, t=0,1, \ldots \tag{5}
\end{equation*}
$$

where $\Gamma=I_{p_{1}}-\sum_{i=1}^{\infty} \Gamma_{i}$ and $\left\|C_{i}\right\|=O\left(\rho^{i}\right)$. Here, $\beta_{\perp}$ is a $p_{1} \times\left(p_{1}-m\right)$ matrix of full rank for which $\beta^{\prime} \beta_{\perp}=0$, and similarly for $\alpha_{\perp}$. This shows that $X_{1 t}$ is a cointegrated $I(1)$ process, that is, $X_{1 t}$ is nonstationary, while $\beta^{\prime} X_{1 t}$ and $\Delta X_{1 t}$ are stationary.

A statistical analysis, including estimation of $\alpha, \beta$, and $\Gamma$, can be conducted for the observations $X_{1 t}, t=1, \ldots T$, using an approximating finite order CVAR, see Saikkonen (1992) and Saikkonen and Lütkepohl (1996).

Hoover (2018) investigates, in particular, whether weak exogeneity for $\beta$ in the approximating finite order CVAR, that is, a zero row in $\alpha$, is a useful tool for finding the causal structure in the graph.

The present note solves the problem of finding expressions for the parameters $\alpha$ and $\beta$ in the $\operatorname{CVAR}(\infty)$ model (4) for the observation $X_{1 t}$, as functions of the parameters in model (3), and finds conditions on these for the presence of a zero row in $\alpha$, and hence weak exogeneity for $\beta$ in the approximating finite order CVAR.

## 2. The Assumptions and Main Results

First, some definitions and assumptions are given, then the main results on $\alpha$ and $\beta$ are presented and proved in Theorems 1 and 2. These results rely on Theorem A1 on the solution of an algebraic Riccati equation, which is given and proved in the Appendix A.

In the following, a $k \times k$ matrix is called stable, if all eigenvalues are contained in the open unit disk. If $A$ is a $k_{1} \times k_{2}$ matrix of rank $k \leq \min \left(k_{1}, k_{2}\right)$, an orthogonal complement, $A_{\perp}$, is defined as a $k_{1} \times\left(k_{1}-k\right)$ matrix of rank $k_{1}-k$ for which $A_{\perp}^{\prime} A=0$. If $k_{1}=k, A_{\perp}=0$. Note that $A_{\perp}$ is only defined up to multiplication from the right by a $\left(k_{1}-k\right) \times\left(k_{1}-k\right)$ matrix of full rank. Throughout, $E_{t}($.$) and$ $\operatorname{Var}_{t}($.$) denote conditional expectation and variance given the sigma-field \mathcal{F}_{0, t}=\sigma\left\{X_{1, s}, 0 \leq s \leq t\right\}$, generated by the observations.

Assumption 1. In Equation (3), it is assumed that
(i) $\varepsilon_{1 t}, \varepsilon_{2 t}$, and $\eta_{t}$ are mutually independent and i.i.d. Gaussian with mean zero and variances $\Omega_{1}, \Omega_{2}$, and $\Omega_{\eta}$, where $\Omega_{1}$ and $\Omega_{2}$ are diagonal matrices,
(ii) $I_{p_{1}}+M_{11}, I_{p_{2}}+M_{22}$ and $I_{p}+M$ are stable,
(iii) $C_{1.2}=C_{1}-M_{12} M_{22}^{-1} C_{2}$ has full rank $m$.

Let $\left(X_{1 t}^{\prime} ; X_{2 t}^{\prime} ; T_{t}^{\prime}\right)^{\prime}, 0=1, \ldots, n$, be the solution to (3) given in (2), such that $\Delta X_{t}$ and $M X_{t}+C T_{t}$ are stationary.

Assumption 1(ii) on $M_{11}, M_{22}$ and $M$ is taken from Hoover (2018) to ensure that, for instance, the process $X_{t}$ given by the equations $X_{t}=\left(I_{p}+M\right) X_{t-1}+$ input, is stationary if the input is stationary, such that the nonstationarity of $X_{t}$ in model (3) is created by the trends $T_{t}$, and not by the own dynamics of $X_{t}$ as given by $M$. It follows from this assumption that $M$ is nonsingular, because $I_{p}+M$ is stable, and similarly for $M_{11}$ and $M_{22}$. Moreover $M_{11.2}=M_{11}-M_{12} M_{22}^{-1} M_{21}$ is nonsingular because

$$
\operatorname{det} M=\operatorname{det} M_{22} \operatorname{det} M_{11.2} \neq 0
$$

## The Main Results

The first result on $\beta$ is a simple consequence of model (3).
Theorem 1. Assumption 1 implies that the cointegrating rank is $r=p_{1}-m$, and that the coefficients $\beta$ and $\beta_{\perp}$ in the $\operatorname{CVAR}(\infty)$ representation for $X_{1 t}$, see (4), are given for $p_{1}>m$ as

$$
\begin{equation*}
\beta_{\perp}=M_{11.2}^{-1} C_{1.2} \text { and } \beta=M_{11.2}^{\prime}\left(C_{1.2}\right)_{\perp} . \tag{6}
\end{equation*}
$$

For $p_{1}=m, \beta_{\perp}$ has rank $p_{1}$, and there is no cointegration: $\alpha=\beta=0$.
Proof of Theorem of 1. From the model Equation (3), it follows, by eliminating $X_{2 t}$ from the first two equations, that

$$
\Delta X_{1, t+1}-M_{12} M_{22}^{-1} \Delta X_{2, t+1}=M_{11.2} X_{1 t}+C_{1.2} T_{t}+\varepsilon_{1 t+1}-M_{12} M_{22}^{-1} \varepsilon_{2, t+1}
$$

Solving for the nonstationary terms gives

$$
\begin{equation*}
M_{11.2} X_{1 t}+C_{1.2} T_{t}=\Delta X_{1, t+1}-M_{12} M_{22}^{-1} \Delta X_{2, t+1}-\varepsilon_{1 t+1,}+M_{12} M_{22}^{-1} \varepsilon_{2, t+1} \tag{7}
\end{equation*}
$$

Multiplying by $\beta^{\prime} M_{11.2}^{-1}$, it is seen that $\beta^{\prime} X_{1 t}$ is stationary, if $\beta^{\prime} M_{11.2}^{-1} C_{1.2}=0$. By Assumption 1(i), $C_{1.2}$ has rank $m$, so that $\beta$ has rank $p_{1}-m$, which proves (6).

The result for $\alpha$ is more involved and is given in Theorem 2. The proof is a further analysis of (7) and involves first, the representation $X_{1 t}$ in terms of a sum of prediction errors $v_{t}^{\beta}=\Delta X_{1 t}-$ $E\left(\Delta X_{1 t} \mid \mathcal{F}_{t-1}^{\beta}\right)$, see (5), and second, a representation of $E\left(T_{t} \mid \mathcal{F}_{0, t}\right)=E\left(T_{t} \mid X_{10}, \ldots, X_{1 t}\right)$ as the (weighted) sum of the prediction errors $v_{0 t}=\Delta X_{1 t}-E\left(\Delta X_{1 t} \mid \mathcal{F}_{0, t-1}\right)$. The second representation requires a result from control theory on the solution of an algebraic Riccati equation, together with some results based on the Kalman filter for the calculation of the conditional mean and variance of the unobserved processes $X_{2 t}, T_{t}$ given the observations $X_{0 s}, 0 \leq s \leq t$. These are collected as Theorem A1 in the Appendix A.

For the discussion of these results, it is useful to reformulate (3) by defining the unobserved variables and errors

$$
T_{t}^{*}=\binom{X_{2 t}}{T_{t}}, \eta_{t}^{*}=\binom{\varepsilon_{2 t}}{\eta_{t}}, \Omega^{*}=\operatorname{Var}\left(\eta_{t}^{*}\right)=\left(\begin{array}{cc}
\Omega_{2} & 0  \tag{8}\\
0 & \Omega_{\eta}
\end{array}\right)
$$

and the matrices

$$
Q^{*}=\left(\begin{array}{cc}
I_{p_{2}}+M_{22} & C_{2}  \tag{9}\\
0 & I_{m}
\end{array}\right), M_{21}^{*}=\binom{M_{21}}{0}, C^{*}=\left(M_{12} ; C_{1}\right)
$$

Then, (3) becomes

$$
\begin{align*}
& X_{1, t+1}=\left(I_{p_{1}}+M_{11}\right) X_{1 t}+C^{*} T_{t}^{*}+\varepsilon_{1, t+1},  \tag{10}\\
& T_{t+1}^{*}=M_{21}^{*} X_{1 t}+Q^{*} T_{t}^{*}+\eta_{t+1}^{*} .
\end{align*}
$$

One can then show, see Theorem A1, that based on properties of the Gaussian distribution, a recursion can be found for the calculation of $V_{t}=\operatorname{Var}_{t}\left(T_{t}^{*}\right)$ and $E_{t}=E_{t}\left(T_{t}^{*}\right)=E_{t}\left(T_{t}^{*} \mid \mathcal{F}_{0 t}\right)$ and $V_{t}=\operatorname{Var}_{t}\left(T_{t}^{*}\right)=\operatorname{Var}_{t}\left(T_{t}^{*} \mid \mathcal{F}_{0 t}\right)$, using the matrices in (8) and (9), by the equations Some

$$
\begin{align*}
& V_{t+1}=Q^{*} V_{t} Q^{* \prime}+\Omega^{*}-Q^{*} V_{t} C^{* \prime}\left(C^{*} V_{t} C^{* \prime}+\Omega_{1}\right)^{-1} C^{*} V_{t} Q^{* \prime},  \tag{11}\\
& E_{t+1}=M_{21}^{*} X_{1 t}+Q^{*} E_{t}+Q^{*} V_{t} C^{* \prime}\left(C^{*} V_{t} C^{* \prime}+\Omega_{1}\right)^{-1} v_{0 t+1} . \tag{12}
\end{align*}
$$

It then follows from results from control theory, that $V=\lim _{t \rightarrow \infty} \operatorname{Var}_{t}\left(T_{t}^{*}\right)$ exists and satisfies the algebraic Riccati equation

$$
\begin{equation*}
V=Q^{*} V Q^{* \prime}+\Omega^{*}-Q^{*} V C^{* \prime}\left(C^{*} V C^{* \prime}+\Omega_{1}\right)^{-1} C^{*} V Q^{* \prime} \tag{13}
\end{equation*}
$$

Moreover, the prediction errors $v_{0 t}=\Delta X_{1 t}-E\left(\Delta X_{1 t} \mid \mathcal{F}_{0, t-1}\right)$ are independent $N_{p_{1}}\left(0, \Sigma_{t}\right)$ for $\Sigma_{t}=C^{*} V_{t} C^{* \prime}+\Omega_{1}$, and the prediction errors $v_{t}^{\beta}=\Delta X_{1 t}-E\left(\Delta X_{1 t} \mid \mathcal{F}_{t-1}^{\beta}\right)$ are independent identically distributed $N_{p_{1}}(0, \Sigma)$ for $\Sigma=C^{*} V C^{* \prime}+\Omega_{1}$. Finally, $E_{t}\left(T_{t}\right)$ has the representation in the prediction errors, $v_{0 i}$,

$$
\begin{equation*}
E_{t}\left(T_{t}\right)=E_{0}\left(T_{0}\right)+\left(0 ; I_{m}\right) \sum_{i=1}^{t} V_{i} C^{* \prime} \Sigma_{i}^{-1} v_{0 i} \tag{14}
\end{equation*}
$$

where $E_{0}\left(T_{0}\right)=E\left(T_{0} \mid X_{10}\right)=0$.

Comparing the representation (5) for $X_{1 t}$ and (14) for $E_{t}\left(T_{t}\right)$ gives a more precise relation between the coefficients of the nonstationary terms in (7). The main result of the paper is to show how this leads to expressions for the coefficients $\alpha$ and $\alpha_{\perp}$ as functions of the parameters in model (3).

Theorem 2. Assumption 1 implies, that the coefficients $\alpha$ and $\alpha_{\perp}$ in the $\operatorname{CVAR}(\infty)$ representation of $X_{1 t}$ are given for $p_{1}>m$ as

$$
\begin{equation*}
\alpha_{\perp}=\Sigma^{-1}\left(M_{12} V_{2 T}+C_{1} V_{T T}\right), \alpha=\Sigma\left(M_{12} V_{2 T}+C_{1} V_{T T}\right)_{\perp}, \tag{15}
\end{equation*}
$$

where

$$
\Sigma=\operatorname{Var}\left(v_{t}^{\beta}\right)=C^{*} V C^{* \prime}+\Omega_{1}=\left(M_{12} ; C_{1}\right)\left(\begin{array}{cc}
V_{22} & V_{2 T}  \tag{16}\\
V_{T 2} & V_{T T}
\end{array}\right)\left(M_{12} ; C_{1}\right)^{\prime}+\Omega_{1} .
$$

Proof of Theorem 2. The left hand side of (7) has two nonstationary terms. The observation $X_{1 t}$ is represented in (5) in terms of a random walk in the prediction errors $v_{i}^{\beta}$, plus a stationary term, and $T_{t}$ is a random walk in $\eta_{i}$. Calculating the conditional expectation given the sigma-field $\mathcal{F}_{0, t}, T_{t}$ is replaced by $E_{t}\left(T_{t}\right)$, which in (14) is represented as a weighted sum of $v_{0 i}$. Thus, the conditional expectation of (7) gives

$$
\begin{equation*}
M_{11.2} X_{1 t}+C_{1.2} E_{t}\left(T_{t}\right)=E_{t}\left(\Delta X_{1 t+1}-M_{12} M_{22}^{-1} \Delta X_{2, t+1}\right), \tag{17}
\end{equation*}
$$

where the right hand side is bounded in mean:

$$
E\left|E_{t}\left(\Delta X_{1, t+1}-M_{12} M_{22}^{-1} \Delta X_{2, t+1}\right)\right| \leq c\left\{E\left|\Delta X_{1, t+1}\right|+\left|\Delta X_{2, t+1}\right|\right\} \leq c
$$

Setting $t=[n u]$ and dividing by $n^{1 / 2}$, it follows from (5) that

$$
\begin{equation*}
n^{-1 / 2} X_{1[n u]} \xrightarrow{\mathrm{D}} \beta_{\perp}\left(\alpha_{\perp}^{\prime} \Gamma \beta_{\perp}\right)^{-1} \alpha_{\perp}^{\prime} W_{v}(u), \tag{18}
\end{equation*}
$$

where $W_{v}(u)$ is the Brownian motion generated by the i.i.d. prediction errors $v_{t}^{\beta}$.
From (14), it can be proved that

$$
\begin{equation*}
n^{-1 / 2} E_{[n u]}\left(T_{[n u]}\right)=\left(0 ; I_{m}\right) n^{-1 / 2} \sum_{t=1}^{[n u]} V_{t} C^{* \prime} \Sigma_{t}^{-1} v_{0 t} \xrightarrow{\mathrm{D}}\left(0 ; I_{m}\right) V C^{* \prime} \Sigma^{-1} W_{v}(u) . \tag{19}
\end{equation*}
$$

This follows by replacing $V_{t}, \Sigma_{t}$ by $V, \Sigma$, because for $\delta_{t}^{\prime}=V_{t} C^{* \prime} \Sigma_{t}^{-1}-V C^{* \prime} \Sigma^{-1} \rightarrow 0$, it holds that

$$
\operatorname{Var}\left(n^{-1 / 2} \sum_{t=1}^{[n u]} \delta_{t}^{\prime} \nu_{0 t}\right)=n^{-1} \sum_{t=1}^{[n u]} \delta_{t}^{\prime} \Sigma_{t} \delta_{t} \rightarrow 0, n \rightarrow \infty .
$$

Next we can replace $v_{0 t}$ by $v_{t}^{\beta}$ as follows: For $t=0,1, \ldots$ the sum

$$
\alpha \beta^{\prime} X_{1 t}+\sum_{i=1}^{t} \Gamma_{i} \Delta X_{1, t+1-i}=\alpha \beta^{\prime} X_{1 t}+\Gamma_{1} \Delta X_{1 t}+\cdots+\Gamma_{t} \Delta X_{11}
$$

is measurable with respect to both $\mathcal{F}_{t}^{\beta}$ and $\mathcal{F}_{0 t}$, such that

$$
v_{0, t+1}-v_{t+1}^{\beta}=-E\left(\sum_{i=t+1}^{\infty} \Gamma_{i} \Delta X_{1, t+1-i} \mid \mathcal{F}_{0, t}\right)+\sum_{i=t+1}^{\infty} \Gamma_{i} \Delta X_{1, t+1-i} .
$$

Then

$$
E\left|v_{0, t+1}-v_{t+1}^{\beta}\right| \leq c \sum_{i=t+1}^{\infty} \rho^{i} E\left|\Delta X_{1, t+1-i}\right|=O\left(\rho^{t}\right)
$$

and therefore

$$
E\left|n^{-1 / 2} \sum_{i=1}^{[n u]}\left(v_{t+1}^{\beta}-v_{0, t+1}\right)\right| \leq n^{-1 / 2} \sum_{i=1}^{[n u]} E\left|v_{t+1}^{\beta}-v_{0, t+1}\right| \leq c n^{-1 / 2} \sum_{i=1}^{[n u]} \rho^{i} \rightarrow 0, n \rightarrow \infty,
$$

which proves (19).
Finally, setting $t=[n u]$ and normalizing (17) by $n^{-1 / 2}$, it follows that in the limit

$$
M_{11.2} \beta_{\perp}\left(\alpha_{\perp}^{\prime} \Gamma \beta_{\perp}\right)^{-1} \alpha_{\perp}^{\prime} W_{v}(u)+C_{1.2}\left(0 ; I_{m}\right) V C^{* \prime} \Sigma^{-1} W_{v}(u)=0 \text { for } u \in[0,1]
$$

This relation shows that the coefficient to $W_{v}(u)$ is zero, so that $\alpha_{\perp}$ can be chosen as

$$
\alpha_{\perp}=\Sigma^{-1} C^{*} V\left(0 ; I_{m}\right)^{\prime}=\Sigma^{-1}\left(M_{12} V_{2 T}+C_{1} V_{T T}\right)
$$

and therefore $\alpha=\Sigma\left(M_{12} V_{2 T}+C_{1} V_{T T}\right)_{\perp}$ which proves (15).

## 3. Two Examples of Simplifying Assumptions

It follows from Theorem 2 that in order to investigate a zero row in $\alpha$, the matrix $V$ is needed. This is easy to calculate from the recursion (11), for a given value of the parameters, but the properties of $V$ are more difficult to evaluate. In general, $\alpha$ does not contain a zero row, but if $M_{12} V_{2 T}=0$, the expressions for $\alpha$ and $\alpha_{\perp}$ simplify, so that simple conditions on $M_{12}$ and $C_{1}$ imply a zero row in $\alpha$ and hence give weak exogeneity in the statistical analysis of the approximating finite order CVAR. This extra condition, $M_{12} V_{2 T}=0$, implies that

$$
\Sigma=\left(M_{12} ; C_{1}\right) V\left(M_{12} ; C_{1}\right)^{\prime}+\Omega_{1}=M_{12} V_{22} M_{12}^{\prime}+C_{1} V_{T T} C_{1}^{\prime}+\Omega_{1}
$$

and

$$
\left(M_{12} V_{2 T}+C_{1} V_{T T}\right)_{\perp}=\left(C_{1} V_{T T}\right)_{\perp}=C_{1 \perp}
$$

such that $\alpha$ simplifies to

$$
\alpha=\left(M_{12} V_{22} M_{12}^{\prime}+C_{1} V_{T T} C_{1}^{\prime}+\Omega_{1}\right) C_{1 \perp}=\left(M_{12} V_{22} M_{12}^{\prime}+\Omega_{1}\right) C_{1 \perp}
$$

Thus, a condition for a zero row in $\alpha$ is

$$
\begin{equation*}
e_{i}^{\prime} \alpha=e_{i}^{\prime} M_{12} V_{22} M_{12}^{\prime} C_{1 \perp}+\omega_{i} e_{i}^{\prime} C_{1 \perp}=0 \tag{20}
\end{equation*}
$$

because $\Omega_{1}=\operatorname{diag}\left(\omega_{1}, \ldots, \omega_{p_{1}}\right)$. This is simple to check by inspecting the matrices $M_{12}$ and $C_{1 \perp}$ in model (3). In the next section, two cases are given, where such a simple solution is available.

Case $1\left(M_{12}=0\right)$. If the unobserved process $X_{2 t}$ does not cause the observation $X_{1 t}$, then $M_{12}=0$. Therefore, $M_{12} V_{2 T}=0$ and from (20) it follows that

$$
e_{i}^{\prime} \alpha=\omega_{i} e_{i}^{\prime} C_{1 \perp}=0
$$

Thus, $\alpha$ has a zero row if $C_{1 \perp}$ has a zero row.
An example of $M_{12}=0$ is the chain $T \rightarrow x_{1} \rightarrow x_{2} \rightarrow x_{3}$, where $X_{1}=\left\{x_{1}, x_{2}, x_{3}\right\}$ is observed and $X_{2}=0$, and hence $M_{12}=0$ and $C_{2}=0$. Then, because $T \rightarrow x_{1}$

$$
C_{1}=\left(\begin{array}{c}
* \\
0 \\
0
\end{array}\right), C_{1 \perp}=\left(\begin{array}{ll}
0 & 0 \\
1 & 0 \\
0 & 1
\end{array}\right)
$$

Thus, the first row of $C_{1 \perp}$ is a zero row, such that $x_{1}$ is weakly exogenous.
To formulate the next case, a definition of strong orthogonality of two matrices is introduced.
Definition 1. Let $A$ be $a k \times k_{1}$ matrix and $B a k \times k_{2}$ matrix. Then, $A$ and $B$ are called strongly orthogonal if $A^{\prime} D B=0$ for all diagonal matrices $D$, or equivalently if $A_{j i} B_{j \ell}=0$ for all $i, j, \ell$.

Thus, if $A_{j i} \neq 0$, we assume that row $j$ of $B$ is zero, and if $B_{j \ell} \neq 0$, row $j$ of $A$ is zero. A simple example is

$$
A=\left(\begin{array}{cc}
* & * \\
0 & * \\
0 & 0
\end{array}\right), B=\left(\begin{array}{c}
0 \\
0 \\
*
\end{array}\right)
$$

Thus, the definition means that if two matrices are strongly orthogonal, it is due to the positions of the zeros and not to linear combination of nonzero numbers being zero.

Thus, in particular if $M_{12}$ and $C_{1}$ are strongly orthogonal, and if $T$ causes a variable in $X_{1}$, then $X_{2}$ does not cause that variable. The expression for $V$ simplifies in the following case.

Lemma 1. If $C_{2}=0$, and $M_{12}^{\prime} \Omega_{1}^{-1} C_{1}=0$, then $Q^{*}=\operatorname{blockdiag}\left(I_{p_{2}}+M_{22} ; I_{m}\right)$, and $V_{2 T}=0$ such that $V=\operatorname{blockdiag}\left(V_{22} ; V_{T T}\right)$.

Proof of Lemma 1. We first prove that $V_{t}$ is blockdiagonal for $t=0$. From (2), it follows that

$$
\binom{X_{10}}{X_{20}}=M^{-1} \sum_{i=0}^{\infty}\left(I_{p}+M\right)^{i}\left(M \varepsilon_{-i}+C \eta_{-i}\right) \text { and } T_{0}=0 .
$$

Thus, if $\Phi$ denotes the variance of $\left(X_{10}^{\prime} ; X_{20}^{\prime}\right)^{\prime}$, then

$$
V_{0}=\operatorname{Var}\left(\left.\binom{X_{20}}{T_{0}} \right\rvert\, X_{10}\right)=\left(\begin{array}{cc}
\Phi_{22.1} & 0 \\
0 & 0
\end{array}\right)
$$

and hence blockdiagonal. Assume, therefore, that $V_{t}=\operatorname{blockdiag}\left(V_{t 22} ; V_{t T T}\right)$ and consider the expression for $V_{t+1}$, see (11). In this expression, $Q^{*}$ is block diagonal (because $C_{2}=0$ ) and $Q^{*} V_{t} Q^{* \prime}$ and $\Omega^{*}$ are block diagonal, and the same holds for $Q^{*} V_{t}^{1 / 2}$. Thus, it is enough to show that

$$
V_{t}^{1 / 2} C^{* \prime}\left\{C^{*} V_{t} C^{* \prime}+\Omega_{1}\right\}^{-1} C^{*} V_{t}^{1 / 2}
$$

is block diagonal. To simplify the notation, define the normalized matrices

$$
\check{M}=\Omega_{1}^{-1 / 2} M_{12} V_{t 22}^{1 / 2} \text { and } \check{C}=\Omega_{1}^{-1 / 2} C_{1} V_{t T T}^{1 / 2}
$$

Then, by assumption,

$$
\check{M}^{\prime} \check{C}=V_{t 22}^{1 / 2} M_{12}^{\prime} \Omega_{1}^{-1} C_{1} V_{t T T}^{1 / 2}=0
$$

so that, using $V_{t 2 T}=0$,

$$
V_{t}^{1 / 2} C^{* \prime}\left(C^{*} V_{t} C^{* \prime}+\Omega_{1}\right)^{-1} C^{*} V_{t}^{1 / 2}=(\check{M}, \check{C})^{\prime}\left(\check{M} \check{M}^{\prime}+\check{C} \check{C}^{\prime}+I_{p_{1}}\right)^{-1}(\check{M}, \check{C})
$$

A direct calculation shows that

$$
\left(\check{M} \check{M}^{\prime}+\check{C} \check{C}^{\prime}+I_{p_{1}}\right)^{-1}=I_{p_{1}}-\check{M}\left(I_{p_{2}}+\check{M}^{\prime} \check{M}\right)^{-1} \check{M}^{\prime}-\check{C}\left(I_{p_{2}}+\check{C}^{\prime} \check{C}\right)^{-1} \check{C}^{\prime}
$$

and that

$$
\check{M}^{\prime}\left\{I_{p_{1}}-\check{M}\left(I_{p_{2}}+\check{M}^{\prime} \check{M}\right)^{-1} \check{M}^{\prime}-\check{C}\left(I_{p_{2}}+\check{C}^{\prime} \check{C}\right)^{-1} \check{C}^{\prime}\right\} \check{C}=0
$$

such that $(\check{M}, \check{C})^{\prime}\left(\check{M} \check{M}^{\prime}+\check{C} \check{C}^{\prime}+I_{p_{1}}\right)^{-1}(\check{M}, \check{C})$ is block diagonal.
Then, $V_{t}^{1 / 2} C^{* \prime}\left\{C^{*} V_{t} C^{* \prime}+\Omega_{1}\right\}^{-1} C^{*} V_{t}^{1 / 2}$ and hence $V_{t+1}$ are block diagonal. Taking the limit for $t \rightarrow \infty$, it is seen that also $V$ is block diagonal.

Case 2 ( $C_{2}=0$, and $M_{12}$ and $C_{1}$ are strongly orthogonal). Because $C_{2}=0$ and $M_{21}^{\prime} \Omega_{1}^{-1} C_{1}=0$, Lemma 1 shows that $V_{2 T}=0$, so that the condition $M_{12} V_{2 T}=0$ and (20) hold. Moreover, strong orthogonality also implies that $M_{12}^{\prime} C_{1}=0$ such that $M_{12}=C_{1 \perp \xi}$ for some $\xi$. Hence

$$
\begin{equation*}
e_{i}^{\prime} \alpha=e_{i}^{\prime} M_{12} V_{22} M_{12}^{\prime} C_{1 \perp}+\omega_{i} e_{i}^{\prime} C_{1 \perp}=e_{i}^{\prime} C_{1 \perp}\left(\xi V_{22} M_{12}^{\prime} C_{1 \perp}+\omega_{i} I_{p_{1}-m}\right), \tag{21}
\end{equation*}
$$

and therefore, a zero row in $\mathrm{C}_{1 \perp}$ gives a zero row in $\alpha$.
Consider again the chain $T \rightarrow x_{1} \rightarrow x_{2} \rightarrow x_{3}$, but assume now that $x_{2}$ is not observed. Thus, $X_{1}=$ $\left\{x_{1}, x_{3}\right\}$ and $X_{2}=\left\{x_{2}\right\}$. Here, $T$ causes $x_{1}$, and $x_{2}$ causes $x_{3}$, so that

$$
M_{12}=\binom{0}{*}, C_{1}=\binom{*}{0}, C_{2}=0 .
$$

Note that $M_{12}^{\prime} D C_{1}=0$ for all diagonal $D$ because $T$ and $X_{2}$ cause disjoint subsets of $X_{1}$. This, together with $C_{2}=0$, implies that $V$ is block diagonal and that (21) holds. Thus, $x_{i}$ is weakly exogenous, $e_{i}^{\prime} \alpha=0$, if

$$
e_{i}^{\prime} C_{1 \perp}=e_{i}^{\prime}\binom{0}{*}=0 .
$$

## 4. Conclusions

This paper investigates the problem of finding adjustment and cointegrating coefficients for the infinite order CVAR representation of a partially observed simple CVAR(1) model. The main tools are some classical results for the solution of the algebraic Riccati equation, and the results are exemplified by an analysis of CVAR(1) models for causal graphs in two cases where simple conditions for weak exogeneity are derived in terms of the parameters of the $\operatorname{CVAR}(1)$ model.

Funding: This research received no external funding
Acknowledgments: The author would like to thank Kevin Hoover for long discussions on the problem and its solution, and Massimo Franchi for reading a first version of the paper and for pointing out the excellent book by Lancaster and Rodman, and two anonymous referees who helped clarify some of the proofs.
Conflicts of Interest: The author declares no conflict of interest.

## Appendix A.

The next Theorem shows how the Kalman filter can be used to calculate $\operatorname{Var}_{t}\left(T_{t}^{*}\right)$ and $E_{t}\left(T_{t}^{*}\right)$ using the same technique as for the common trends model and proves the existence of the limit of $V_{t}$. The last result follows from the theory of the algebraic Riccati equation, see Lancaster and Rodman (1995), in the following $\operatorname{LR}(1995)$.

Theorem A1. Let $X_{1 t}$ and $T_{t}^{*}$ be given by model (10) and let Assumption 1 be satisfied. Then, $V_{t}=\operatorname{Var}_{t}\left(T_{t}^{*}\right)$ and $E_{t}=E_{t}\left(T_{t}^{*}\right)$ are given recursively, using the starting values $E_{0}$ and $V_{0}$ by

$$
\begin{align*}
V_{t+1} & =Q^{*} V_{t} Q^{* \prime}+\Omega^{*}-Q^{*} V_{t} C^{* \prime} \Sigma_{t}^{-1} C^{*} V_{t} Q^{* \prime}  \tag{A1}\\
E_{t+1} & =M_{21}^{*} X_{1 t}+Q^{*} E_{t}+Q^{*} V_{t} C^{* \prime} \Sigma_{t}^{-1} v_{0, t+1} \tag{A2}
\end{align*}
$$

where

$$
\begin{equation*}
\Sigma_{t}=C^{*} V_{t} C^{* \prime}+\Omega_{1}, \tag{A3}
\end{equation*}
$$

and the prediction errors

$$
\begin{equation*}
v_{0, t+1}=X_{1, t+1}-E_{t}\left(X_{1, t+1}\right) \tag{A4}
\end{equation*}
$$

are independent $N_{p_{1}}\left(0, \Sigma_{t}\right)$.
The sequence $V_{t}$ starting with $V_{0}$, converges to a finite positive limit $V$, which satisfies the algebraic Riccati equation,

$$
\begin{equation*}
V=Q^{*} V Q^{* \prime}+\Omega^{*}-Q^{*} V C^{* \prime} \Sigma^{-1} C^{*} V Q^{* \prime}, \quad \Sigma=C^{*} V C^{* \prime}+\Omega_{1} \tag{A5}
\end{equation*}
$$

Furthermore,

$$
\begin{equation*}
Q^{*}-Q^{*} V C^{* /} \Sigma^{-1} C^{*} \tag{A6}
\end{equation*}
$$

is stable, and $E_{t}\left(T_{t}\right)$ satisfies the equation

$$
\begin{equation*}
E_{t+1}\left(T_{t+1}\right)=E_{t}\left(T_{t}\right)+\left(0 ; I_{m}\right) V_{t} C^{* /} \Sigma_{t}^{-1} v_{0, t+1} \tag{A7}
\end{equation*}
$$

Proof of Theorem A1. The variance $V_{t}=\operatorname{Var}_{t}\left(T_{t}^{*}\right)$ can be calculated recursively, using the properties of the Gaussian distribution, as

$$
\begin{align*}
\operatorname{Var}_{t+1}\left(T_{t+1}^{*}\right) & =\operatorname{Var}_{t}\left(T_{t+1}^{*} \mid X_{1, t+1}\right)  \tag{A8}\\
& =\operatorname{Var}_{t}\left(T_{t+1}^{*}\right)-\operatorname{Cov}_{t}\left(T_{t+1}^{*} ; X_{1, t+1}\right) \operatorname{Var}_{t}\left(X_{1, t+1}\right)^{-1} \operatorname{Cov}_{t}\left(X_{1, t+1} ; T_{t+1}^{*}\right) .
\end{align*}
$$

From the model Equation (10), it follows that

$$
\begin{align*}
\operatorname{Var}_{t}\left(T_{t+1}^{*}\right) & =\operatorname{Var}_{t}\left\{M_{21}^{*} X_{1 t}+Q^{*} T_{t}^{*}+\eta_{t+1}^{*}\right\}=Q^{*} \operatorname{Var}_{t}\left(T_{t}^{*}\right) Q^{* \prime}+\Omega^{*},  \tag{A9}\\
\operatorname{Cov}_{t}\left(T_{t+1}^{*} ; X_{1, t+1}\right) & =\operatorname{Cov}_{t}\left\{T_{t+1}^{*} ;\left(I_{p_{1}}+M_{11}\right) X_{1 t}+C^{*} T_{t}^{*}+\varepsilon_{1 t+1}\right\}=Q^{*} \operatorname{Var}_{t}\left(T_{t}^{*}\right) C^{* \prime},  \tag{A10}\\
\operatorname{Var}_{t}\left(X_{1, t+1}\right) & =\operatorname{Var}_{t}\left\{\left(I_{p_{1}}+M_{11}\right) X_{1 t}+C^{*} T_{t}^{*}+\varepsilon_{1 t+1}\right\}=C^{*} \operatorname{Var}_{t}\left(T_{t}^{*}\right) C^{* \prime}+\Omega_{1} . \tag{A11}
\end{align*}
$$

Then, (A8)-(A11) give the recursion for $V_{t}=\operatorname{Var}_{t}\left(T_{t}^{*}\right)$ in (A1). Similarly, for the conditional mean, it is seen that

$$
\begin{aligned}
E_{t+1}\left(T_{t+1}^{*}\right) & =E_{t}\left(T_{t+1}^{*} \mid X_{1, t+1}\right)=E_{t}\left(T_{t+1}^{*}\right)+\operatorname{Cov}_{t}\left(T_{t+1}^{*} ; X_{1, t+1}\right) \operatorname{Var}_{t}\left(X_{1, t+1}\right)^{-1} v_{0, t+1}, \\
E_{t}\left(T_{t+1}^{*}\right) & =M_{21}^{*} X_{1 t}+Q^{*} E_{t}\left(T_{t}^{*}\right)
\end{aligned}
$$

which implies (A2) with prediction error $v_{0, t+1}=\Delta X_{1, t+1}-E_{t}\left(\Delta X_{1, t+1}\right)$.
Note that (A1) is the usual recursion from the Kalman filter equations for the state space model obtained from (10) for $M_{21}^{*}=0$, see Durbin and Koopman (2012). Note also, however, that (A2) is not the usual recursion from the common trends model, because of the first term containing $M_{21}^{*}$. It is seen from (A1) that if $V_{t}$ converges to $V$, then $V$ has to satisfy the algebraic Riccati equation (A5) and $\Sigma$ is given as indicated.

The result that $V_{t}$ converges to a finite positive limit follows from LR (1995, Theorem 17.5.3), where the assumptions, in the present notation, are
a. $1\left(Q^{*} ; I_{p_{2}+m}\right)$ is controllable,
a. $2\left(Q^{*} ; I_{p_{2}+m}\right)$ is stabilizable,
a. $3\left(C^{*} ; Q^{*}\right)$ is detectable.

Before giving the proof, some definitions from control theory are given, which are needed for checking the conditions of the results in $\operatorname{LR}(1995)$.

Let $A$ be a $k \times k$ matrix and $B$ be a $k \times k_{1}$ matrix.
d. 1 The pair $\{A, B\}$ is called controllable if

$$
\operatorname{rank}\left(B ; A B ; \ldots ; A^{k-1} B\right)=k
$$

LR(1995, (4.1.3)).
d. 2 The pair $\{A ; B\}$ is stabilizable if there is a $k_{1} \times k$ matrix $K$, such that $A+B K$ is stable $\operatorname{LR}(1995$, page 90 , line $5-$ ).
d. 3 Finally $\{B ; A\}$ is detectable means that $\left\{A^{\prime} ; B^{\prime}\right\}$ is stabilizable, $\operatorname{LR}(1995$, page 91 line 6-).

The first assumption, $a .1$, is easy to check: The pair $\left(Q^{*} ; I_{p_{2}+m}\right)$ is controllable, see $d .1$, means that

$$
\operatorname{rank}\left(I_{p_{2}+m} ; Q^{*} I_{p_{2}+m} ; \ldots ; Q^{* p_{2}+m-1} I_{p_{2}+m}\right)=p_{2}+m
$$

The second assumption, a.2, follows because controllability implies stabilizability, see LR (1995, Theorem 4.4.2).

Finally, d. 3 shows that $\left(C^{*} ; Q^{*}\right)$ detectable means $\left(Q^{* \prime} ; C^{* \prime}\right)$ stabilizable, and $\operatorname{LR}(1995$, Theorem 4.5.6 (b)), see also Hautus (1969), shows that $\left(Q^{* \prime} ; C^{* \prime}\right)$ is stabilizable, if and only if

$$
\operatorname{rank}\left(Q^{* \prime}-\lambda I_{p_{2}+m} ; C^{* \prime}\right)=\operatorname{rank}\left(\begin{array}{cc}
M_{12} & C_{1} \\
I_{p_{2}}+M_{22}-\lambda I_{p_{2}} & C_{2} \\
0 & I_{m}-\lambda I_{m}
\end{array}\right)=p_{2}+m \text { for all }|\lambda| \geq 1
$$

For $\lambda=1$, using $C_{1.2}=C_{1}-M_{12} M_{22}^{-1} C_{2}$ and Assumption 1, it follows that

$$
\begin{aligned}
\operatorname{rank}(M(1)) & =\operatorname{rank}\left(\begin{array}{ll}
M_{12} & C_{1} \\
M_{22} & C_{2}
\end{array}\right)=\operatorname{rank}\left(\begin{array}{cc}
0 & C_{1.2} \\
M_{22} & C_{2}
\end{array}\right) \\
& =\operatorname{rank}\left(C_{1.2}\right)+\operatorname{rank}\left(M_{22}\right)=m+p_{2} .
\end{aligned}
$$

For $|\lambda|>1$, using Assumption 1(ii), it is seen that

$$
\operatorname{rank}(M(\lambda))=\operatorname{rank}\left(I_{p_{2}}+M_{22}-\lambda I_{p_{2}}\right)+\operatorname{rank}\left(I_{m}-\lambda I_{m}\right)=p_{2}+m
$$

because $\lambda$ is not an eigenvalue of the stable matrix $I_{p_{2}}+M_{22}$, when $|\lambda|>1$.
Thus, $\left(Q^{* \prime} ; C^{* \prime}\right)$ is stabilizable, and assumptions $a .1, a .2, a .3$ hold, such that and LR (1995, Theorem 17.5.3) applies. This proves that limit $V=\lim _{t \rightarrow \infty} V_{t}$ exists and (A6) holds.

Multiplying (A2) by $\left(0 ; I_{m}\right)$, it is seen, using $\left(0 ; I_{m}\right) Q^{*}=\left(0 ; I_{m}\right)$, and $\left(0 ; I_{m}\right) M_{21}^{*}=0$, that a recursion for $E_{t}\left(T_{t}\right)$ is given by (A7).

## References

Durbin, James, and Siem Jan Koopman. 2012. Time Series Analysis by State Space Methods, 2nd ed. Oxford: Oxford University Press.
Hautus, Malo L. J. 1969. Controllability and observability conditions of linear autonomous systems. Koninklijke Nederlandse Akademie van Wetenschappen. Indagationes Mathematicae 12: 443-48.
Hoover, Kevin D. 2018. Long-Run Causal Order: A Preliminary Investigation. Durham: Department of Economics and Department of Philosophy, Duke University.
Johansen, Søren, and Katarina Juselius. 2014. An asymptotic invariance property of the common trends under linear transformations of the data. Journal of Econometrics 17: 310-15. [CrossRef]
Lancaster, Peter, and Leiba Rodman. 1995. Algebraic Riccati Equations. Oxford: Clarendon Press.
Saikkonen, Pentti. 1992. Estimation and testing of cointegrated systems by an autoregressive approximation. Econometric Theory 8: 1-27. [CrossRef]
Saikkonen, Pentti, and Helmut Lütkepohl. 1996. Infinite order cointegrated vector autoregressive processes. Estimation and Inference. Econometric [CrossRef] Theory 12: 814-44.

Article

# Cointegration and Error Correction Mechanisms for Singular Stochastic Vectors 

Matteo Barigozzi ${ }^{1}$, Marco Lippi ${ }^{2, *}$ and Matteo Luciani ${ }^{3}$<br>1 Università di Bologna, Department of Economics, 40126 Bologna, Italy; matteo.barigozzi@unibo.it<br>2 Einaudi Institute for Economics and Finance, 00187 Roma, Italy<br>3 Federal Reserve Board of Governors, Washington, DC 20551, USA; matteo.luciani@frb.gov<br>* Correspondence: mlippi.eief@gmail.com

Received: 28 March 2018; Accepted: 7 January 2020; Published: 4 February 2020


#### Abstract

Large-dimensional dynamic factor models and dynamic stochastic general equilibrium models, both widely used in empirical macroeconomics, deal with singular stochastic vectors, i.e., vectors of dimension $r$ which are driven by a $q$-dimensional white noise, with $q<r$. The present paper studies cointegration and error correction representations for an $I(1)$ singular stochastic vector $\mathbf{y}_{t}$. It is easily seen that $\mathbf{y}_{t}$ is necessarily cointegrated with cointegrating rank $c \geq r-q$. Our contributions are: (i) we generalize Johansen's proof of the Granger representation theorem to $I(1)$ singular vectors under the assumption that $\mathbf{y}_{t}$ has rational spectral density; (ii) using recent results on singular vectors by Anderson and Deistler, we prove that for generic values of the parameters the autoregressive representation of $\mathbf{y}_{t}$ has a finite-degree polynomial. The relationship between the cointegration of the factors and the cointegration of the observable variables in a large-dimensional factor model is also discussed.


Keywords: singular stochastic vectors; cointegration for singular vectors; Granger representation theorem; large-dimensional dynamic factor models)

JEL Classification: C0; C01; E0

## 1. Introduction

An $r$-dimensional stochastic vector $\mathbf{y}_{t}$ such that $\mathbf{y}_{t}=\mathbf{A}_{0} \mathbf{u}_{t}+\mathbf{A}_{1} \mathbf{u}_{t-1}+\cdots$, where the matrices $\mathbf{A}_{j}$ are $r \times q$ and $\mathbf{u}_{t}$ is a $q$-dimensional white noise, with $q<r$, is said to be singular. Singular stochastic vectors have been systematically analyzed in a number of papers starting with (Anderson and Deistler 2008a, 2008b). A motivation for studying the consequences of singularity, as argued by these authors, is that the factors' vector in large-dimensional dynamic factor models (DFM), such as those introduced in Forni et al. (2000); Forni and Lippi (2001), (Stock and Watson 2002a, 2002b), is typically singular. Singularity is also an important feature of dynamic stochastic general equilibrium models (DSGE), see e.g., Sargent (1989), Canova (2007), pp. 230-2. Singularity as it arises in DFMs is presented in some detail below.

DFMs are based on the idea that all the observed variables in an economic system are driven by a few common (macroeconomic) shocks and by idiosyncratic components which may result from measurement errors and sectoral or regional shocks. Formally, each variable in the $n$-dimensional dataset $x_{i t}, i=1,2, \ldots, n, t=1,2, \ldots, T$, is decomposed into the sum of a common component $\chi_{i t}$, and an idiosyncratic component $\epsilon_{i t}: x_{i t}=\chi_{i t}+\epsilon_{i t}$, where $\chi_{i t}$ and $\epsilon_{j s}$ are orthogonal for all $i, j, t, s$. In the standard version of the DFM the common components are linear combinations of an $r$-dimensional vector of common factors $\mathbf{F}_{t}=\left(F_{1 t} F_{2 t} \cdots F_{r t}\right)^{\prime}$,

$$
\begin{equation*}
\chi_{i t}=\lambda_{i 1} F_{1 t}+\lambda_{i 2} F_{2 t}+\cdots+\lambda_{i r} F_{r t}=\lambda_{i} \mathbf{F}_{t} . \tag{1}
\end{equation*}
$$

Now suppose that the observable variables $x_{i t}$ and the common factors $\mathbf{F}_{t}$ are $I(1)$ and that

$$
\begin{equation*}
(1-L) \mathbf{F}_{t}=\mathbf{C}(L) \mathbf{u}_{t}, \tag{2}
\end{equation*}
$$

where $\mathbf{u}_{t}$ is a nonsingular $q$-dimensional white-noise vector ${ }^{1}$, the common shocks. A number of papers analyzing macroeconomic databases find strong empirical support for the assumption that the vector $\mathbf{F}_{t}$ is singular, i.e., that $q<r$. See, for US datasets, Giannone et al. (2005); Amengual and Watson (2007); Forni and Gambetti (2010), Luciani (2015). For a Euro-area dataset, see Barigozzi et al. (2014).

Such results can be easily understood observing that usually the static Equation (1) is just a convenient representation derived from a "primitive" set of dynamic equations linking the common components $\chi_{i t}$ to the common shocks $\mathbf{u}_{t}$. As a simple example, suppose that the variables $x_{i t}$ are driven by a common one-dimensional cyclical process $f_{t}$, such that $(1-\alpha L) f_{t}=u_{t}$, where $u_{t}$ is scalar white noise, and that the variables $x_{i t}$ load $f_{t}$ dynamically:

$$
\begin{equation*}
x_{i t}=a_{i 0} f_{t}+a_{i 1} f_{t-1}+\epsilon_{i t} . \tag{3}
\end{equation*}
$$

In this case we can set $F_{1 t}=f_{t}, F_{2 t}=f_{t-1}=F_{1, t-1}, \lambda_{i 1}=a_{i 0}, \lambda_{i 2}=a_{i 1}$, so that Equations (1) and (2) take the form

$$
x_{i t}=\lambda_{i 1} F_{1 t}+\lambda_{i 2} F_{2 t}+\epsilon_{i t} \text { and }\binom{F_{1 t}}{F_{2 t}}=\binom{(1-\alpha L)^{-1}}{L(1-\alpha L)^{-1}} u_{t}
$$

respectively. Here $r=2$ and $q=1$ so that $\mathbf{F}_{t}$ is singular. For a general analysis of the relationship between representation (1) and "deeper" dynamic representations like (3), see e.g., Forni et al. (2009); Stock and Watson (2016).

Now suppose that the factors $\mathbf{F}_{t}$ have been estimated. Obtaining $\mathbf{u}_{t}$ and the impulse-response functions of the variables $x_{i t}$ with respect to $\mathbf{u}_{t}$ (or structural shocks obtained by a linear transformation of $\mathbf{u}_{t}$ ) requires the estimation of a VAR for the singular $I(1)$ vector $\mathbf{F}_{t}$. On the other hand, the latter is necessarily cointegrated with cointegration rank $c$ at least equal to $r-q$ (the rank of the spectral density of $(1-L) \mathbf{F}_{t}$ does not exceed $q$ at all frequencies and, therefore, at frequency zero).

Singular vectors of factors in an $I(1)$ DFM and $I(1)$ singular vectors in DSGE models provide strong motivation for studying singular $I(1)$ vectors in a general time-series context. The main contributions of the paper are:
(I) A generalization of Johansen's proof of the Granger Representation Theorem (from MA to AR), this is Proposition 2. Consider an $I(1)$ singular vector $\mathbf{y}_{t}$, with dimension $r$, rank $q<r$, and cointegrating rank $c \geq r-q$. Assuming that $(1-L) \mathbf{y}_{t}$ has an ARMA structure, $\mathbf{S}(L)(1-$ $L) \mathbf{y}_{t}=\mathbf{B}(L) \mathbf{u}_{t}$ and that some simple additional conditions hold, $\mathbf{y}_{t}$ has a representation as a vector error correction mechanism (VECM) with c error correction terms:

$$
\begin{equation*}
\mathbf{A}(L) \mathbf{y}_{t}=\mathbf{A}^{*}(L)(1-L) \mathbf{y}_{t}+\boldsymbol{\alpha}\left(\boldsymbol{\beta}^{\prime} \mathbf{y}_{t-1}-\mathbf{w}\right)=\mathbf{B}(0) \mathbf{u}_{t} \tag{4}
\end{equation*}
$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are both $r \times c$ and full rank, $\boldsymbol{\beta}^{\prime} \mathbf{y}_{t}-\mathbf{w}$ is $I(0), \mathbf{A}(L)$ and $\mathbf{A}^{*}(L)$ are $r \times r$ rational matrices in $L$. Under the additional assumption that unity is the only zero of $\mathbf{B}(L)$, i.e., if $z \neq 1$ then $\mathbf{B}(z)$ is full rank, $\mathbf{A}(L)$ and $\mathbf{A}^{*}(L)$ are finite-degree matrix polynomials.
(II) Assuming that the parameters of $\mathbf{S}(L)$ and $\mathbf{B}(L)$ may vary in an open subset of $\mathbb{R}^{\lambda}$, see Section 3.2 for the definition of $\lambda$, in Proposition 3 we show that all the assumptions used to obtain (4), and also the assumption that unity is the only possible zero of $\mathbf{B}(L)$, hold for generic values of

[^11]the parameters. This implies that the matrices $\mathbf{A}(L)$ and $\mathbf{A}^{*}(L)$ are generically of finite degree, which is obviously not the case for nonsingular vectors. ${ }^{2}$

The paper is organized as follows. Section 2 is preliminary. We firstly recall recent results for stationary singular stochastic vectors with rational spectral density, see (Anderson and Deistler 2008a, 2008b). Secondly, we discuss cointegration and the cointegrating rank for $I(1)$ singular stochastic vectors.

In Section 3 we prove our main results. We also obtain the permanent-transitory shock representation in the singular case: $\mathbf{y}_{t}$ is driven by $r-c$ permanent shocks, i.e., $r$ minus the cointegrating rank, the usual result. However, the number of transitory shocks is $c-(r-q)$, not $c$ as in the nonsingular case.

Section 3 also contains an exercise carried out with simulated singular $I(1)$ vectors. We compare the results obtained by estimating an unrestricted VAR in the levels and a VECM. Though limited to a simple example, the results confirm what has been found for nonsingular vectors, that under cointegration the long-run features of impulse-response functions are better estimated using a VECM rather than an unrestricted VAR in the levels (Phillips 1998).

In Section 4 we analyse cointegration of the observable variables $x_{i t}$ in a DFM. Our results on cointegration of the singular vector $\mathbf{F}_{t}$ have the implication that $p$-dimensional subvectors of the $n$-dimensional common-component vector $\chi_{t}$, with $p>r-c$, are cointegrated. As a consequence, stationarity of the idiosyncratic components would imply that all $p$-dimensional subvectors of the $n$-dimensional dataset $\mathbf{x}_{t}$ are cointegrated if $p>r-c$. For example, if $q=3$ and $d=1$, then all 3-dimensional subvectors in the dataset are cointegrated, a kind of regularity that we do not observe in actual large macroeconomic datasets. This suggests that an estimation strategy robust to the assumption that the idiosyncratic components can be $I(1)$ has to be preferred (for this aspect we refer to Barigozzi et al. 2019). Section 5 concludes. Some proofs, a discussion of some non-uniqueness problems arising with singularity and details on the simulations are collected in the Appendix.

## 2. Stationary and I(1) Singular Vectors

### 2.1. Stationary Singular Vectors

As in this paper we only consider representation issues it is convenient to assume that all stochastic processes are defined for $t \in \mathbb{Z}$. Accordingly, the lag operator $L$ is defined as $L \mathbf{y}_{t}=\mathbf{y}_{t-1}$ for $t \in \mathbb{Z}$ (Bauer and Wagner (2012) also study $I(1)$ and cointegrated processes for $t \in \mathbb{Z}$ ).

We start by introducing results on singular vectors with an ARMA structure from (Anderson and Deistler 2008a, 2008b). Some preliminary definitions are needed.

## Definition 1. (Zeros and Poles)

(A) When considering matrices $\mathbf{V}(z)$ whose entries are rational functions of $z \in \mathbb{C}$ we always assume that numerator and denominator of each entry have no common roots. If $\mathbf{V}(z)$ is an $r \times q$ matrix of rational functions, we say that $z^{*}$ is a pole of $\mathbf{V}(z)$ if it is a pole of some entry of $\mathbf{V}(z)$.
(B) Suppose that $\mathbf{V}(z)$ is an $r \times q$ matrix whose entries are polynomial functions of $z \in \mathbb{C}$, with $q \leq r$. We say that $z^{*} \in \mathbb{C}$ is a zero of $\mathbf{V}(z)$ if $\operatorname{rank}\left(\mathbf{V}\left(z^{*}\right)\right)<q$, and that $\mathbf{V}(z)$ is zeroless if it has no zeros, i.e., $\operatorname{rank}(\mathbf{V}(z))=q$ for all $z \in \mathbb{C}$.

[^12]With a minor abuse of language, we may speak of zeros and poles of the corresponding matrix $\mathbf{V}(L)$. When a $r \times r$ polynomial matrix $\mathbf{S}(L)$ has all its zeros outside the unit circle we say that $\mathbf{S}(L)$ is stable.

All the stationary vector processes considered have an ARMA structure. Precisely, the $r$-dimensional process $\mathbf{y}_{t}$ has an ARMA structure with rank $q, q \leq r$, if there exist
(i) a non-singular $q$-dimensional white-noise process $\mathbf{u}_{t}$,
(ii) an $r \times r$ stable polynomial matrix $\mathbf{S}(z)$, with $\mathbf{S}(0)=\mathbf{I}_{r}$,
(iii) an $r \times q$ matrix $\mathbf{B}(z)$ whose rank is $q$ for all $z$ with the exception of a finite subset of $\mathbb{C}$, such that

$$
\begin{equation*}
\mathbf{y}_{t}=\mathbf{V}(L) \mathbf{u}_{t} \tag{5}
\end{equation*}
$$

where $\mathbf{V}(L)=\mathbf{S}(L)^{-1} \mathbf{B}(L)$.
Suppose that $\mathbf{y}_{t}$ has also the representation $\mathbf{y}_{t}=\tilde{\mathbf{S}}(L)^{-1} \tilde{\mathbf{B}}(L) \tilde{\mathbf{u}}_{t}$, where $\tilde{\mathbf{u}}_{t}$ is a $\tilde{q}$-dimensional nonsingular white noise. Denoting by $\boldsymbol{\Sigma}_{y}(\theta)$ the spectral density of $\mathbf{y}_{t}$,

$$
\boldsymbol{\Sigma}_{y}(\theta)=(2 \pi)^{-1} \mathbf{V}\left(e^{-i \theta}\right) \boldsymbol{\Sigma}_{u} \mathbf{V}^{\prime}\left(e^{i \theta}\right)
$$

so that the rank of $\boldsymbol{\Sigma}_{y}(\theta)$ is $q$ for all $\theta$, with the exception of a finite subset of $[-\pi, \pi]$. As the spectral density is independent of the ARMA representation, $q=\tilde{q}$ and $\tilde{\mathbf{B}}(z)$ has rank $q$ except for a finite subset of $\mathbb{C}$.

Remark 1. Let us recall that the equation

$$
\mathbf{S}(L) \boldsymbol{\tau}_{t}=\mathbf{B}(L) \mathbf{u}_{t}
$$

in the unknown vector process $\zeta_{t}$, where $\mathbf{S}(L)$ is stable, has only one stationary solution, and this is $\mathbf{y}_{t}=$ $\mathbf{S}(L)^{-1} \mathbf{B}(L) \mathbf{u}_{t}$. Thus the ARMA process $\mathbf{y}_{t}$ can also be defined as the stationary solution of $\mathbf{S}(L) \boldsymbol{\zeta}_{t}=\mathbf{B}(L) \mathbf{u}_{t}$.

Definition 2. (Genericity) Suppose that a statement $Q$ depends on $\mathbf{p} \in \mathcal{A}$, where $\mathcal{A}$ is an open subset of $\mathbb{R}^{\lambda}$. We say that $Q$ holds generically in $\mathcal{A}$, or that $Q$ holds for generic values of $\mathbf{p} \in \mathcal{A}$, if the subset $\mathcal{N}$ of $\mathcal{A}$ where it does not hold is nowhere dense in $\mathcal{A}$, i.e., the closure of $\mathcal{N}$ in $\mathcal{A}$ has no internal points.

For example, assuming that $\mathbf{p} \in \mathcal{A}=\mathbb{R}$, the statement "The roots of the polynomial $x^{2}+\mathbf{p} x+1$ are distinct" holds generically in $\mathcal{A}$.

Definition 3. (Rational reduced-rank family of filters) Assume that $r>q$ and let $\mathcal{G}$ be a set of ordered couples $(\mathbf{S}(L), \mathbf{B}(L))$, where:
(i) $\mathbf{B}(L)$ is an $r \times q$ polynomial matrix of degree $s_{1} \geq 0$.
(ii) $\mathbf{S}(L)$ is an $r \times r$ polynomial matrix of degree $s_{2} \geq 0 . \mathbf{S}(0)=\mathbf{I}_{r}$.
(iii) Denoting by $\mathbf{p}$ the vector containing the $\lambda=r q\left(s_{1}+1\right)+r^{2} s_{2}$ coefficients of the entries of $\mathbf{B}(L)$ and $\mathbf{S}(L)$, we assume that $\mathbf{p} \in \Pi$, where $\Pi$ is an open subset of $\mathbb{R}^{\lambda}$ such that for $\mathbf{p} \in \Pi$, (1) $\mathbf{S}(z)$ is stable, (2) $\operatorname{rank}(\mathbf{B}(z))=q$ with the exception of a finite subset of $\mathbb{C}$.

We say that $\mathcal{G}$ is a rational reduced-rank family of filters with parameter set $\Pi$.
The notation $\mathbf{S}^{\mathbf{p}}(L), \mathbf{B}^{\mathbf{p}}(L)$, though more rigorous, would be heavy and not really necessary. We use it only in Appendix A.1.

Proposition 1. Assume that $r>q$.
(I) Suppose that $\mathbf{V}(L)$ is an $r \times q$ matrix polynomial in $L$. If $\mathbf{V}(z)$ is zeroless then $\mathbf{V}(L)$ has an $r \times r$ finite-degree stable left inverse, i.e., there exists a finite-degree polynomial $r \times r$ matrix $\mathbf{W}(L)$ such that:
(a) $\mathbf{W}(0)=\mathbf{I}_{r}$, (b) $\operatorname{det}(\mathbf{W}(z))=0$ implies $|z|>1$, (c) $\mathbf{W}(L) \mathbf{V}(L)=\mathbf{V}(0)$. Let $\mathbf{y}_{t}$ be the stationary solution of $\mathbf{S}(L) \zeta_{t}=\mathbf{B}(L) \mathbf{u}_{t}$ and suppose that $\mathbf{B}(L)$ is zeroless. Then $\mathbf{y}_{t}$ has a finite vector autoregressive representation (VAR) $\mathbf{A}(L) \mathbf{y}_{t}=\mathbf{B}(0) \mathbf{u}_{t}$, where $\mathbf{A}(L)=\mathbf{N}(L) \mathbf{S}(L)$ and $\mathbf{N}(L)$ is a finite-degree left inverse of $\mathbf{B}(L)$.
(II) Assume that $\mathbf{y}_{t}$ is the stationary solution of $\mathbf{S}(L) \zeta_{t}=\mathbf{B}(L) \mathbf{u}_{t}$, where $(\mathbf{S}(L), \mathbf{B}(L))$ belongs to a rational reduced-rank family of filters with parameter set $\Pi$. For generic values of the parameters in $\Pi, \mathbf{B}(L)$ is zeroless so that $\mathbf{y}_{t}$ has a finite VAR representation.

For statement (I) see Anderson and Deistler (2008a), Theorem 3. Statement (II) is a modified version of their Theorem 2, see for a proof Forni et al. (2009), p. 1327.

### 2.2. Fundamentalness

Assume that the $r$-dimensional vector $\mathbf{y}_{t}$ has an ARMA structure, rank $q$ and the moving average representation (5). If $\operatorname{rank}(\mathbf{B}(z))=q$ for $|z|<1$, then $\mathbf{u}_{t}$ belongs to the space spanned by $\mathbf{y}_{t-k}$, with $k \geq 0$, and representation (5), as well as $\mathbf{u}_{t}$, is called fundamental (for these definitions and results see e.g., Rozanov (1967), pp. 43-7). Note that if (5) is fundamental $\operatorname{rank}(\mathbf{B}(0))=q$. Note also that when $q=r$, the condition that $\operatorname{rank}(\mathbf{B}(z))=q$ for $|z|<1$ becomes $\operatorname{det}(\mathbf{B}(z)) \neq 0$ for $|z|<1$.

Remark 2. Note that in Proposition 1, part (II), we do not assume that $\mathbf{u}_{t}$ is fundamental for $\mathbf{y}_{t}$. However, Proposition 1, (II), states that for generic values of $\mathbf{p} \in \Pi$ the matrix $\mathbf{B}(L)$ is zeroless and therefore $\mathbf{u}_{t}$ is fundamental for $\mathbf{y}_{t}$.

### 2.3. I(1) Singular Vectors

To analyze cointegration and the autoregressive representations of singular non-stationary vectors let us first recall the definitions of $I(0), I(1)$ and cointegrated vectors. This requires some preliminary definitions and results.

We denote by $L_{2}(\Omega, \mathcal{F}, P)$ the space of the square-integrable functions on the probability space $(\Omega, \mathcal{F}, P)$. Let $\mathbf{z}_{t}=\left(z_{1 t} z_{2 t} \cdots z_{r t}\right)^{\prime}, z_{h t} \in L_{2}(\Omega, \mathcal{F}, P)$, be an $r$-dimensional stochastic process and consider the difference equation

$$
\begin{equation*}
(1-L) \zeta_{t}=\mathbf{z}_{t} \tag{6}
\end{equation*}
$$

in the unknown $r$-dimensional process $\boldsymbol{\zeta}_{t}$. A solution of (6) is

$$
\tilde{\psi}_{t}=\left\{\begin{array}{l}
\mathbf{z}_{1}+\mathbf{z}_{2}+\cdots+\mathbf{z}_{t}, \text { for } t>0 \\
\mathbf{0}, \text { for } t=0 \\
-\left(\mathbf{z}_{0}+\mathbf{z}_{-1} \cdots+\mathbf{z}_{t+1}\right), \text { for } t<0
\end{array}\right.
$$

see e.g., Gregoir (1999), p. 439, Franchi and Paruolo (2019). All the solutions of (6) are $\psi_{t}=\tilde{\psi}_{t}+\boldsymbol{\phi}_{t}$, where $\boldsymbol{\phi}_{t}=\left(\phi_{1 t} \phi_{2 t} \cdots \phi_{r t}\right)^{\prime}, \phi_{h t} \in L_{2}(\Omega, \mathcal{F}, P)$, is a solution of the homogeneous equation $(1-L) \boldsymbol{\zeta}_{t}=$ $\mathbf{0}$, so that $\phi_{t}=\mathbf{K}$, for some $r$-dimensional stochastic vector $\mathbf{K}$, for all $t \in \mathbb{Z}$. We say that the process $\boldsymbol{\phi}_{t}=\mathbf{K}$ is a constant stochastic process. Obviously a constant stochastic process $\boldsymbol{\phi}_{t}=\mathbf{K}$ is weakly stationary. Its spectral measure has the jump $\Sigma_{K}$ at frequency zero. Thus $\phi_{t}$ has a spectral density (has an absolutely continuous spectral measure) if and only if $\boldsymbol{\Sigma}_{K}=\mathbf{0}$, i.e., if and only if $\boldsymbol{\phi}_{t}(\omega)=\mathbf{k}$, where $\mathbf{k} \in \mathbb{R}^{r}$, for $\omega$ almost everywhere in $\Omega$.

## Definition 4. (I(0), I(1) and Cointegrated vectors)

$\mathbf{I}(\mathbf{0})$. An $r$-dimensional ARMA $\mathbf{y}_{t}$ with spectral density $\boldsymbol{\Sigma}_{y}(\theta)$ is $I(0)$ if $\boldsymbol{\Sigma}_{y}(0) \neq \mathbf{0}$.
$\mathbf{I}(\mathbf{1})$. The $r$-dimensional vector stochastic process $\mathbf{y}_{t}$ is $I(1)$ if it is a solution $(1-L) \boldsymbol{\zeta}_{t}=\mathbf{z}_{t}$ where $\mathbf{z}_{t}$ is an $r$-dimensional $I(0)$ process. The rank of $\mathbf{y}_{t}$ is defined as the rank of $\mathbf{z}_{t}$.

Cointegration. Assume that the r-dimensional stochastic vector $\mathbf{y}_{t}$ is $I(1)$ and denote by $\boldsymbol{\Sigma}_{\Delta y}(\theta)$ the spectral density of $(1-L) \mathbf{y}_{t}$. The vector $\mathbf{y}_{t}$ is cointegrated with cointegrating rank $c$, with $0<c<r$, if $\operatorname{rank}\left(\boldsymbol{\Sigma}_{\Delta y}(0)\right)=r-c$.

If $q$ is the rank of $\mathbf{y}_{t}$ and $r \geq q$, then $c=r-q+d$, where $q>d>0$. Thus in the singular case, $r>q, \mathbf{y}_{t}$ is necessarily cointegrated with cointegrating rank at least equal to $r-q$.

If $\mathbf{y}_{t}$ is $I(1)$ and cointegrated with cointegrating rank $c$, there exist $c$ linearly independent $r \times 1$ vectors $\mathbf{c}_{j}, j=1, \ldots, c$, such that the spectral density of $\mathbf{c}_{j}^{\prime}(1-L) \mathbf{y}_{t}$ vanishes at frequency zero. The vectors $\mathbf{c}_{j}$ are called cointegrating vectors and the set $\mathbf{c}_{j}, j=1, \ldots, c$, a complete set of cointegrating vectors. Of course a complete set of cointegrating vectors $\mathbf{c}_{j}, j=1, \ldots, c$, can be replaced by the set $\mathbf{d}_{j}$, $j=1, \ldots, c$, where the vectors $\mathbf{d}_{j}$ are $c$ independent linear combinations of the vectors $\mathbf{c}_{j}$.

Lemma 1. (I) Assume that $\mathbf{y}_{t}$ has an ARMA structure and has the rational representation (5): $\mathbf{y}_{t}=\mathbf{V}(L) \mathbf{u}_{t}$. Then $\mathbf{y}_{t}$ is $I(0)$ if and only if $\mathbf{V}(1) \neq \mathbf{0}$.
(II) Assume $(1-L) \mathbf{y}_{t}$ has an ARMA structure and has the rational representation

$$
\begin{equation*}
(1-L) \mathbf{y}_{t}=\mathbf{V}(L) \mathbf{u}_{t} . \tag{7}
\end{equation*}
$$

The process $\mathbf{y}_{t}$ is $I(1)$ if and only if $\mathbf{V}(1) \neq \mathbf{0}$.
(III) If $\mathbf{y}_{t}$ is $I(1)$, cointegrated and has representation (7), the cointegrating rank of $\mathbf{y}_{t}$ is $c$ if and only if the rank of $\mathbf{V}(1)$ is $r-c$. Moreover $\mathbf{c}$ is a cointegrating vector for $\mathbf{y}_{t}$ if and only if $\mathbf{c}^{\prime} \mathbf{V}(1)=\mathbf{0}$.
(IV) Assume that $\mathbf{y}_{t}$ is $I(1)$. c is a cointegrating vector for $\mathbf{y}_{t}$ if and only if a scalar stochastic variable $w \in L_{2}(\Omega, \mathcal{F}, P)$ can be determined such that $\mathbf{c}^{\prime} \mathbf{y}_{t}-w$ is stationary with an ARMA structure.

Proof. (I) is an immediate consequence of $\boldsymbol{\Sigma}_{y}(0)=(2 \pi)^{-1} \mathbf{V}(1) \boldsymbol{\Gamma}_{u} \mathbf{V}(1)^{\prime}$, where $\boldsymbol{\Gamma}_{u}$ is the nonsingular covariance matrix of $\mathbf{u}_{t}$. (II) and (III) are obtained in the same way from $\boldsymbol{\Sigma}_{\Delta y}(0)=(2 \pi)^{-1} \mathbf{V}(1) \boldsymbol{\Gamma}_{u} \mathbf{V}(1)^{\prime}$. (IV) The process $\mathbf{y}_{t}$ solves (6) with $\mathbf{z}_{t}=\mathbf{V}(L) \mathbf{u}_{t}$, so that, defining

$$
\boldsymbol{\mu}_{t}=\left\{\begin{array}{l}
\mathbf{u}_{1}+\mathbf{u}_{2}+\cdots+\mathbf{u}_{t}, \text { for } t>0  \tag{8}\\
\mathbf{0}, \text { for } t=0 \\
-\left(\mathbf{u}_{0}+\mathbf{u}_{-1} \cdots+\mathbf{u}_{t+1}\right), \text { for } t<0
\end{array}\right.
$$

we have

$$
\mathbf{y}_{t}=\mathbf{V}(L) \boldsymbol{\mu}_{t}+\mathbf{K}=\left[\mathbf{V}(1)+(1-L) \frac{\mathbf{V}(L)-\mathbf{V}(1)}{1-L}\right] \boldsymbol{\mu}_{t}+\mathbf{K}=\mathbf{V}(1) \boldsymbol{\mu}_{t}+\mathbf{V}^{*}(L) \mathbf{u}_{t}+\mathbf{K}
$$

where (i) the entries of $\mathbf{V}^{*}(L)=(\mathbf{V}(L)-\mathbf{V}(1)) /(1-L)$ are rational functions of $L$ with no poles of modulus less or equal to unity, (ii) $\mathbf{K}$ is a constant $r$-dimensional stochastic process. We have:

$$
\begin{equation*}
\mathbf{c}^{\prime} \mathbf{y}_{t}=\mathbf{c}^{\prime} \mathbf{V}(1) \boldsymbol{\mu}_{t}+\mathbf{c}^{\prime} \mathbf{V}^{*}(L) \mathbf{u}_{t}+\mathbf{c}^{\prime} \mathbf{K} \tag{9}
\end{equation*}
$$

If $\mathbf{c}$ is a cointegrating vector of $\mathbf{y}_{t}$ we have $\mathbf{c}^{\prime} \mathbf{V}(1)=\mathbf{0}$, so that

$$
\mathbf{c}^{\prime} \mathbf{y}_{t}=\mathbf{c}^{\prime} \mathbf{V}^{*}(L) \mathbf{u}_{t}+\mathbf{c}^{\prime} \mathbf{K}
$$

Setting $w=\mathbf{c}^{\prime} \mathbf{K}$, the process $\mathbf{c}^{\prime} \mathbf{y}_{t}-w=\mathbf{c}^{\prime} \mathbf{V}^{*}(L) \mathbf{u}_{t}$ has the desired properties. Note that $w$ has the equivalent definition $w=\mathbf{c}^{\prime} \mathbf{y}_{0}-\mathbf{c}^{\prime} \mathbf{V}^{*}(L) \mathbf{u}_{0}$. Conversely, suppose that $w$ is such that $\mathbf{c}^{\prime} \mathbf{y}_{t}-w$ has an ARMA structure. By (9),

$$
\mathbf{c}^{\prime} \mathbf{y}_{t}-w=\mathbf{c}^{\prime} \mathbf{V}(1) \boldsymbol{\mu}_{t}+\mathbf{c}^{\prime} \mathbf{V}^{*}(L) \mathbf{u}_{t}+\mathbf{c}^{\prime} \mathbf{K}-w
$$

so that

$$
\sqrt{\mathrm{E}\left(\mathbf{c}^{\prime} \mathbf{y}_{t}-w\right)^{2}}+\sqrt{\mathrm{E}\left(\mathbf{c}^{\prime} \mathbf{V}^{*}(L) \mathbf{u}_{t}\right)^{2}}+\sqrt{\mathrm{E}\left(\mathbf{c}^{\prime} \mathbf{K}-w\right)^{2}} \geq \sqrt{\mathbf{c}^{\prime} \mathbf{V}(1) \boldsymbol{\Sigma}_{\mu_{t}} \mathbf{V}^{\prime}(1) \mathbf{c}}
$$

The three terms on the left-hand side are finite and independent of $t$. As $\boldsymbol{\Sigma}_{\mu_{t}}=|t| \boldsymbol{\Sigma}_{u}$ and $\boldsymbol{\Sigma}_{u}$ is positive definite, the right-hand side diverges for $|t| \rightarrow \infty$ unless $\mathbf{c}^{\prime} \mathbf{V}(1)=\mathbf{0}$.

Lemma 1 shows that our definitions of $I(0)$ and $I(1)$ processes are equivalent to Definitions 3.2, and 3.3 in Johansen (1995), p. 35, with two minor differences: (i) our assumption of rational spectral density, (ii) the time span of the stochastic processes is $t=0,1, \ldots$ in Johansen's book, $t \in \mathbb{Z}$ in the present paper. Also, under the assumption that $(1-L) \mathbf{y}_{t}$ has an ARMA structure, our definition of cointegration is equivalent to that in Johansen (1995), p. 37.

## 3. Representation Theory for Singular I(1) Vectors

In Section 3.1 we prove our generalization to singular vectors of the Granger representation theorem (from MA to AR). We closely follow the proof in Johansen (1995), Theorem 4.5, p. 55-57. In Section 3.2 we show that, under a suitable parameterization, the matrix of the autoregressive representation is generically of finite degree.

### 3.1. The Granger Representation Theorem ( $M A$ to $A R$ )

Suppose that $r \geq q, c>0$ and $r>c \geq r-q$. Let $\mathbf{B}(L)$ be an $r \times q$ polynomial matrix of degree $s_{1} \geq 0$ and $\mathbf{S}(L)$ an $r \times r$ polynomial matrix of degree $s_{2} \geq 0$ with $\mathbf{S}(0)=\mathbf{I}_{r}$.

Assumption 1. $\mathrm{S}(L)$ is stable.
Assumption 2. If $z^{*}$ is a zero of $\mathbf{B}(z)$ (i.e. $\left.\operatorname{rank}\left(\mathbf{B}\left(z^{*}\right)\right)<q\right)$ then either $z^{*}=1$ or $\left|z^{*}\right|>1$.
Assumption 2 implies that the rank of $\mathbf{B}(0)$ is $q$. The next is a stronger version of Assumption 2:
Assumption 3. If $z^{*}$ is a zero of $\mathbf{B}(z)$ then $z^{*}=1$.
Assumption 4. $\operatorname{rank}(\mathbf{B}(1))=r-c$.
Under Assumption 1, let $\mathbf{y}_{t}$ be a solution of the equation

$$
\begin{equation*}
(1-L) \zeta_{t}=\mathbf{S}(L)^{-1} \mathbf{B}(L) \mathbf{u}_{t} \tag{10}
\end{equation*}
$$

We have

$$
\begin{equation*}
\mathbf{y}_{t}=\mathbf{S}(L)^{-1} \mathbf{B}(L) \boldsymbol{\mu}_{t}+\mathbf{K} \tag{11}
\end{equation*}
$$

where $\boldsymbol{\mu}_{t}$ is defined in (8) and $\mathbf{K}$ is a constant stochastic process. By Assumption $4, \mathbf{S}(1)^{-1} \mathbf{B}(1) \neq \mathbf{0}$, so that $\mathbf{y}_{t}$ is $I(1)$ with cointegrating rank $c$, see Lemma 1, (II) and (III).

Consider the finite Taylor expansion of $\mathbf{B}(z)$ around $z=1$ :

$$
\mathbf{B}(z)=\mathbf{B}(1)-(1-z) \mathbf{B}^{\prime}(1)+(1-z)^{2} \mathbf{B}^{\prime \prime}(1)+\cdots
$$

Assumption 4 implies that

$$
\mathbf{B}(1)=\xi \eta^{\prime}
$$

where $\boldsymbol{\xi}$ is $r \times(r-c)$ of rank $r-c, \eta$ is $q \times(r-c)$ of rank $r-c$, see Lancaster and Tismenetsky (1985, p. 97, Proposition 3). The Taylor expansion above can be rewritten as

$$
\begin{equation*}
\mathbf{B}(z)=\xi \eta^{\prime}+(1-z) \mathbf{B}^{*}+(1-z)^{2} \mathbf{E}(z) \tag{12}
\end{equation*}
$$

where $\mathbf{B}^{*}=-\mathbf{B}^{\prime}(1)$ and $\mathbf{E}(z)$ is a polynomial matrix.
Let $\boldsymbol{\xi}_{\perp}$ be an $r \times c$ matrix whose columns are orthogonal to all columns of $\boldsymbol{\xi}$ : (i) the columns of $\xi_{\perp}$ are a complete set of cointegrating vectors for $\mathbf{B}(L) \mathbf{u}_{t}$, (ii) the columns of the matrix $\mathbf{S}^{\prime}(1) \xi_{\perp}$ are a complete set of cointegrating vectors for $\mathbf{y}_{t}$. Regarding (i), using (11) and (12), we have

$$
\begin{equation*}
\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{S}(L) \mathbf{y}_{t}=\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{B}(L) \boldsymbol{\mu}_{t}+\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{S}(1) \mathbf{K}=\left(\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{B}^{*}+(1-L) \boldsymbol{\xi}_{\perp}^{\prime} \mathbf{E}(L)\right) \mathbf{u}_{t}+\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{S}(1) \mathbf{K}, \tag{13}
\end{equation*}
$$

so that $\xi_{\perp}^{\prime} \mathbf{S}(L) \mathbf{y}_{t}-\xi_{\perp}^{\prime} \mathbf{S}(1) \mathbf{K}$ has an ARMA structure. Regarding (ii), see the proof of Proposition 2.
Assumption 5. rank $\left[\binom{\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{B}^{*}}{\boldsymbol{\eta}^{\prime}}\right]=\operatorname{rank}\left[\binom{\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{B}^{*}}{\boldsymbol{\xi}^{\prime} \boldsymbol{\xi} \eta^{\prime}}\right]=q$.
Define $\mathbf{S}^{*}(L)=\frac{\mathbf{S}(L)-\mathbf{S}(1)}{1-L}$.
Assumption 6. $\boldsymbol{\xi}_{\perp}^{\prime}\left(\mathbf{B}^{*}-\mathbf{S}^{*}(1) \mathbf{S}(1)^{-1} \boldsymbol{\xi} \eta^{\prime}\right) \neq \mathbf{0}$.
Remark 3. Let $\mathbf{y}_{t}$ be a solution of (10) so that $(1-L) \mathbf{y}_{t}$ is stationary and $\mathbf{S}(L)\left[(1-L) \mathbf{y}_{t}\right]=\mathbf{B}(L) \mathbf{u}_{t}$. Assumption 2, and therefore 3, implies that $\mathbf{u}_{t}$ is fundamental for $(1-L) \mathbf{y}_{t}$, see Section 2.2.

We are now ready for our main representation result.
Proposition 2. (I) Weak form. Suppose that Assumptions 1, 2, 4, 5 and 6 hold and let $\mathbf{y}_{t}$ be a solution of the difference Equation (10), so that $\mathbf{y}_{t}=\mathbf{S}(L)^{-1} \mathbf{B}(L) \boldsymbol{\mu}_{t}+\mathbf{K}$, with $\boldsymbol{\mu}_{t}$ defined in (8) and $\mathbf{K}$ a constant stochastic process. Set $\boldsymbol{\beta}=\mathbf{S}(1)^{\prime} \xi_{\perp}$. Then a $c$-dimensional stochastic vector $\mathbf{w}$ can be determined such that (i) $\boldsymbol{\beta}^{\prime} \mathbf{y}_{t}-\mathbf{w}$ is $I(0)$, (ii) $\mathbf{y}_{t}$ has the error correction representation

$$
\begin{equation*}
\mathbf{A}(L) \mathbf{y}_{t}=\mathbf{A}^{*}(L)(1-L) \mathbf{y}_{t}+\boldsymbol{\alpha}\left(\boldsymbol{\beta}^{\prime} \mathbf{y}_{t-1}-\mathbf{w}\right)=\mathbf{B}(0) \mathbf{u}_{t}, \tag{14}
\end{equation*}
$$

where $\mathbf{A}(L)$ is a rational $r \times r$ matrix with no poles in or on the unit circle, $\mathbf{A}(1)=\mathbf{I}_{r}, \mathbf{A}^{*}(L)=(\mathbf{A}(L)-$ $\mathbf{A}(1) L)(1-L)^{-1}, \boldsymbol{\alpha}$ is $r \times c$ and full rank, $\boldsymbol{\alpha} \boldsymbol{\beta}^{\prime}=\mathbf{A}(1)$.
(II) Strong form. Under Assumptions 1, 3, 4, 5 and 6, statement (I) holds with an $r \times r$ stable, finite-degree matrix polynomial $\mathbf{A}(L)$.

Proof. Multiply both sides of $(1-L) \mathbf{S}(L) \mathbf{y}_{t}=\mathbf{B}(L) \mathbf{u}_{t}$ by the $r \times r$ invertible matrix $\Xi=\binom{\boldsymbol{\xi}_{\perp}^{\prime}}{\boldsymbol{\xi}^{\prime}}$. We obtain

$$
\begin{align*}
& (1-L) \boldsymbol{\Xi} \mathbf{S}(L) \mathbf{y}_{t}=\boldsymbol{\Xi} \mathbf{B}(L) \mathbf{u}_{t} \\
& =\left\{\binom{\mathbf{0}_{c \times q}}{\boldsymbol{\xi}^{\prime} \xi^{\prime} \eta^{\prime}}+(1-L)\binom{\xi_{\perp}^{\prime} \mathbf{B}^{*}}{\xi^{\prime} \mathbf{B}^{*}}+(1-L)^{2}\binom{\xi_{\perp}^{\prime} \mathbf{E}(L)}{\xi^{\prime} \mathbf{E}(L)}\right\} \mathbf{u}_{t}  \tag{15}\\
& =\left(\begin{array}{cc}
(1-L) \mathbf{I}_{c} & \mathbf{0} \\
\mathbf{0} & \mathbf{I}_{r-c}
\end{array}\right)\left\{\binom{\xi_{\perp}^{\prime} \mathbf{B}^{*}}{\boldsymbol{\xi}^{\prime} \xi^{\prime} \eta^{\prime}}+(1-L)\left(\begin{array}{c}
\xi^{\prime} \\
\perp \mathbf{E}(L) \\
\xi^{\prime} \mathbf{B}^{*}
\end{array}\right)+(1-L)^{2}\binom{\mathbf{0}_{c \times q}}{\xi^{\prime} \mathbf{E}(L)}\right\} \mathbf{u}_{t .} .
\end{align*}
$$

Taking the first $c$ rows in (15),

$$
(1-L) \boldsymbol{\xi}_{\perp}^{\prime} \mathbf{S}(L) \mathbf{y}_{t}=(1-L)\left(\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{B}^{*}+(1-L) \boldsymbol{\xi}_{\perp}^{\prime} \mathbf{E}(L)\right) \mathbf{u}_{t} .
$$

This implies that

$$
\begin{equation*}
\xi_{\perp}^{\prime} \mathbf{S}(L) \mathbf{y}_{t}=\left(\tilde{\xi}_{\perp}^{\prime} \mathbf{B}^{*}+(1-L) \xi_{\perp}^{\prime} \mathbf{E}(L)\right) \mathbf{u}_{t}+\mathbf{w} \tag{16}
\end{equation*}
$$

where $\mathbf{w}$ is a $c$-dimensional constant stochastic vector. Comparing with (13), $\mathbf{w}=\xi_{\perp}^{\prime} \mathbf{S}(1) \mathbf{K}$. On the other hand,

$$
\begin{align*}
\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{S}(1) \mathbf{y}_{t}-\mathbf{w} & =\left(\xi_{\perp}^{\prime} \mathbf{S}(L) \mathbf{y}_{t}-\mathbf{w}\right)-\left(\tilde{\xi}_{\perp}^{\prime} \mathbf{S}(L) \mathbf{y}_{t}-\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{S}(1) \mathbf{y}_{t}\right) \\
& =\left(\xi_{\perp}^{\prime} \mathbf{S}(L) \mathbf{y}_{t}-\mathbf{w}\right)-\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{S}^{*}(L)(1-L) \mathbf{y}_{t} \\
& =\left(\xi_{\perp}^{\prime} \mathbf{S}(L) \mathbf{y}_{t}-\mathbf{w}\right)-\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{S}^{*}(L) \mathbf{S}(L)^{-1} \mathbf{B}(L) \mathbf{u}_{t}  \tag{17}\\
& =\left\{\boldsymbol{\xi}_{\perp}^{\prime}\left(\mathbf{B}^{*}-\mathbf{S}^{*}(1) \mathbf{S}(1)^{-1} \boldsymbol{\xi}^{\prime}\right)+(1-L) \mathcal{H}(L)\right\} \mathbf{u}_{t},
\end{align*}
$$

where the last equality has been obtained using (16) and $\mathcal{H}(L)$ is a suitable polynomial matrix. Thus $\boldsymbol{\beta}^{\prime} \mathbf{y}_{t}-\mathbf{w}=\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{S}(1) \mathbf{y}_{t}-\mathbf{w}$ has an ARMA structure. Moreover, by Assumption 6, $\boldsymbol{\beta}^{\prime} \mathbf{y}_{t}-\mathbf{w}$ is $I(0)$.

Joining (16) with the last $r-c$ rows of (15),

$$
\left(\begin{array}{cc}
\mathbf{I}_{c} & \mathbf{0}  \tag{18}\\
\mathbf{0} & (1-L) \mathbf{I}_{r-c}
\end{array}\right) \Xi \mathbf{\Xi} \mathbf{S}(L) \mathbf{y}_{t}-\binom{\mathbf{I}_{c}}{\mathbf{0}_{(r-c) \times c}} \mathbf{w}=\mathbf{M}(L) \mathbf{u}_{t},
$$

where

$$
\begin{equation*}
\mathbf{M}(L)=\binom{\tilde{\xi}_{\perp}^{\prime} \mathbf{B}^{*}}{\boldsymbol{\xi}^{\prime} \xi \eta^{\prime}}+(1-L)\binom{\xi^{\prime}{ }_{\perp} \mathbf{E}(L)}{\boldsymbol{\xi}^{\prime} \mathbf{B}^{*}}+(1-L)^{2}\binom{\mathbf{0}_{c \times q}}{\xi^{\prime} \mathbf{E}(L)} . \tag{19}
\end{equation*}
$$

By (15) and (19),

$$
\mathbf{B}(L)=\boldsymbol{\Xi}^{-1}\left(\begin{array}{cc}
(1-L) \mathbf{I}_{c} & \mathbf{0}  \tag{20}\\
\mathbf{0} & \mathbf{I}_{r-c}
\end{array}\right) \mathbf{M}(L) \cdot{ }^{3}
$$

By Assumption 5, $\mathbf{M}(z)$ has no zero at $z=1$, see (19). On the other hand, (i) if $z^{*}$ is a zero of $\mathbf{M}(z)$ then $z^{*}$ is a zero of $\mathbf{B}(z)$, (ii) if $z^{*}$ is a zero of $\mathbf{B}(z), z^{*} \neq 1$, then $z^{*}$ is a zero of $\mathbf{M}(z)$. Therefore, Assumption 3 implies that $\mathbf{M}(z)$ is zeroless and viceversa. Under Assumption 2, the zeros of $\mathbf{M}(z)$ lie outside the unit circle. In order to conclude the proof we need inverting $\mathbf{M}(L)$ in (18).
(I) Under Assumption 3, Proposition 1, part (I), states that there exists an $r \times r$ stable, finite-degree polynomial matrix $\mathbf{N}(L)=\mathbf{I}_{r}+\mathbf{N}_{1} L+\cdots+\mathbf{N}_{p} L^{p}$, for some $p$, such that: (i) $\mathbf{N}(0)=\mathbf{I}_{r}$, (ii) $\mathbf{N}(L) \mathbf{M}(L)=\mathbf{M}(0)$.
(II) Under Assumption 2, by a standard procedure we remove all the zeros of $\mathbf{M}(z)$ which lie outside the unit circle ${ }^{4}$, then use Proposition 1, part (I), to left-invert the residual zeroless polynomial, thus obtaining an $r \times r$ rational matrix $\mathbf{N}(L)$ such that (i) $\mathbf{N}(L)$ has no poles in or on the unit circle (possible poles of $\mathbf{N}(L)$ are the zeros of $\mathbf{M}(L)$, which lie outside the unit circle), (ii) $\mathbf{N}(0)=\mathbf{I}_{r}$, (iii) $\mathbf{N}(L) \mathbf{M}(L)=\mathbf{M}(0)$. See also Deistler et al. (2010).

Defining

$$
\mathbf{A}(L)=\Xi^{-1} \mathbf{N}(L)\left(\begin{array}{cc}
\mathbf{I}_{c} & \mathbf{0} \\
\mathbf{0} & (1-L) \mathbf{I}_{r-c}
\end{array}\right) \boldsymbol{\Xi} \mathbf{S}(L)=\Xi^{-1} \mathbf{N}(L)\binom{\boldsymbol{\xi}_{\perp}^{\prime}}{(1-L) \boldsymbol{\xi}^{\prime}} \mathbf{S}(L)
$$

and using $\mathbf{M}(0)=\boldsymbol{\Xi} \mathbf{B}(0)$, we have

$$
\mathbf{A}(L) \mathbf{y}_{t}-\boldsymbol{\Xi}^{-1} \mathbf{N}(1)\binom{\mathbf{I}_{c}}{\mathbf{0}_{(r-c) \times c}} \mathbf{w}=\mathbf{B}(0) \mathbf{u}_{t},
$$

[^13]with $\mathbf{A}(0)=\mathbf{I}_{r}$. Defining $\mathbf{A}^{*}(L)=(\mathbf{A}(L)-\mathbf{A}(1) L)(1-L)^{-1}$,
$$
\mathbf{A}^{*}(L)(1-L) \mathbf{y}_{t}+\mathbf{A}(1) \mathbf{y}_{t-1}-\Xi^{-1} \mathbf{N}(1)\binom{\mathbf{I}_{c}}{\mathbf{0}_{(r-c) \times c}} \mathbf{w}=\mathbf{B}(0) \mathbf{u}_{t}
$$

Defining

$$
\boldsymbol{\alpha}=\Xi^{-1} \mathbf{N}(1)\binom{\mathbf{I}_{c}}{\mathbf{0}_{(r-c) \times c}}
$$

we see that $\mathbf{A}(1)=\boldsymbol{\alpha} \boldsymbol{\beta}^{\prime}$ and

$$
\mathbf{A}^{*}(L)(1-L) \mathbf{y}_{t}+\boldsymbol{\alpha}\left(\boldsymbol{\beta}^{\prime} \mathbf{y}_{t-1}-\mathbf{w}\right)=\mathbf{B}(0) \mathbf{u}_{t}
$$

Some remarks are in order.

Remark 4. (I) Under our assumption of an ARMA structure, Assumption 1 corresponds to Definition 3.1 in Johansen's book, see p. 34. Assumption 2 is Johansen's Assumption 1 (see p. 14), adapted for singularity. Assumption 3 has no counterpart in Johansen's nonsingular framework. In Section 3.2 we show that under the parameterization adopted in Definition 5, Assumption 3 holds generically.
(II) Simplifying the model by taking $\mathbf{S}(L)=\mathbf{I}_{r}$, Assumption 5 generalizes to the singular case Johansen's assumption that $\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{C}^{*} \boldsymbol{\eta}_{\perp}$ is full rank (see Theorem 4.5, p. 55; $\mathbf{C}^{*}$ corresponds to our $\mathbf{B}^{*}$ ). For, assuming that $r=q$, multiplying the matrix in Assumption 5 by the nonsingular matrix $\left(\eta_{\perp} \quad \eta\right)$, we obtain that Assumption 5 holds if and only if $\boldsymbol{\zeta}_{\perp}^{\prime} \mathbf{B}^{*} \eta_{\perp}$ is full rank. Assumption 5 is used in the proof of Proposition 2 to invert the matrix $\mathbf{M}(L)$, which remains on the right-hand side after the removal of the unit roots, see Equation (18), which is the same rôle played by Johansen's assumption in his proof.
(III) Under $\mathbf{S}(L)=\mathbf{I}_{r}$, assumption 6 simplifies to $\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{B}^{*} \neq \mathbf{0}$. If $d>0$ Assumption 6 is a consequence of Assumption 5. For, if $d>0$ then $r-c=q-d<q$. On the other hand, $r-c$ is the number of rows of $\boldsymbol{\eta}^{\prime}$, so that Assumption 5 holds only if Assumption 6 holds. In particular, if $r=q$ and $c=d>0$, Assumption 6 is redundant. However if $r>q$ and $d=0$, so that the rank of $\boldsymbol{\eta}^{\prime}$ is $q$, then Assumption 5 holds even if $\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{B}^{*}=\mathbf{0}$. Assumption 6 is necessary in Proposition 2 to prove that the error correction term is $I(0)$, not only stationary.

Remark 5. Uniqueness issues arise with autoregressive representations of singular vectors. For example, suppose that $c=r-q>0$, so that $d=0$. Representation (14) has an $(r-q)$-dimensional error correction term $\boldsymbol{\beta}^{\prime} \mathbf{y}_{t}-\mathbf{w}$. On the other hand, in this case $\mathbf{B}(1)$ has full rank $q$, so that Proposition 1 (I) applies and, in spite of cointegration, $\mathbf{y}_{t}$ has an autoregressive representation in differences

$$
\mathbf{D}(L) \mathbf{S}(L)(1-L) \mathbf{y}_{t}=\mathbf{B}(0) \mathbf{u}_{t}
$$

In Appendix B. 1 we sketch a proof of the statement that in general, $\mathbf{y}_{t}$ has VECM representations with a number of error correction terms ranging from d to c. However, as we show in Appendix B.2, different autoregressive representations of $\mathbf{y}_{t}$ produce the same impulse-response functions. Both in this and the companion paper Barigozzi et al. (2019) the number of error correction terms in the error correction representation for reduced-rank $I(1)$ vectors is always the maximum $c$. It is worth reporting that, in our experiments with simulated data, the best results in estimation of singular VECMs are obtained using $c$ as the number of error correction terms.

Remark 6. Assume for simplicity that $\mathbf{S}(L)=\mathbf{I}_{r}$. From equation (17):

$$
\mathbf{e}_{t}=\boldsymbol{\beta}^{\prime} \mathbf{y}_{t}-\mathbf{w}=\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{y}_{t}-\mathbf{w}=\left\{\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{B}^{*}+(1-L) \mathcal{H}(L)\right\} \mathbf{u}_{t} .
$$

If $r=q$, Assumption 5 implies that $\xi_{\perp}^{\prime} \mathbf{B}^{*}$ has rank $c$, so that no $c$-dimensional vector $\mathbf{d} \neq \mathbf{0}$ can be determined such that some of the coordinates of $\mathbf{d e}_{t}$ is stationary but not $I(0)$. Thus, according to the definition introduced in Franchi and Paruolo (2019), p. 1181, the error term $\mathbf{e}_{t}$ is a "non-cointegrated $I(0)$ process." When $r>q$ and $c \leq q$, i.e., $r \leq 2 q-d$, elementary examples can be produced in which $\mathbf{e}_{t}$ is an $I(0)$ but not a non-cointegrated $I(0)$ process (one is given in Appendix A.2). Thus Assumption 6 only implies that $\mathbf{e}_{t}$ is $I(0)$. Of course, under $c \leq q$, the assumption that $\boldsymbol{\xi}_{\perp}^{\prime}\left(\mathbf{B}^{*}-\mathbf{S}^{*}(1) \mathbf{S}(1)^{-1} \boldsymbol{\xi} \eta^{\prime}\right)$ has rank $c$, an enhancement of Assumption 6, implies that $\mathbf{e}_{t}$ is a non-cointegrated $I(0)$ process. On the other hand, if $c>q$, i.e., $r>2 q-d, \mathbf{e}_{t}$ cannot be a non-cointegrated $I(0)$ process.

### 3.2. Generically, $\mathbf{A}(L)$ Is a Finite-Degree Polynomial

Suppose that the couple $(\mathbf{S}(L), \mathbf{B}(L))$ is parameterized as in Definition 3. It easy to see that $\mathbf{B}(1)$ has generically rank $q$, so that generically the cointegrating rank of $\mathbf{y}_{t}$ is $r-q$. In particular, if $r=q$ cointegration is non generic.

It is quite easy to see that this paradoxical result only depends on the choice of a parameter set that is unfit to study cointegration. Our starting point here is that a specific value of $c$ between $r-q$ and $r-1$ has a motivation in economic theory or in statistical inference, and must be therefore built in the parameter set. Thus in Definition 5 below the family of filters is redefined so that generically the cointegrating rank is equal to a given $c$ between $r-q$ and $r-1$.

Definition 5. (Rational reduced-rank family of filters with cointegrating rank c) Assume that $r>q$, $c>0$ and $r>c \geq r-q$. Let $\mathcal{G}$ be a set of couples $(\mathbf{S}(L), \mathbf{B}(L))$, where:
(i) The matrix $\mathbf{B}(L)$ has the parameterization

$$
\mathbf{B}(L)=\xi \eta^{\prime}+(1-L) \mathbf{B}^{*}+(1-L)^{2} \mathbf{E}(L),
$$

where $\boldsymbol{\xi}$ and $\eta$ are $r \times(r-c)$ and $q \times(r-c)$ respectively, $\mathbf{B}^{*}$ is an $r \times q$ matrix and $\mathbf{E}(L)$ is an $r \times q$ matrix polynomial of degree $s_{1} \geq 0$.
(ii) $\mathbf{S}(L)$ is an $r \times r$ polynomial matrix of degree $s_{2} \geq 0 . \mathbf{S}(0)=\mathbf{I}_{r}$.
(iii) Denoting by $\mathbf{p}$ the vector containing the $\lambda=(r-c)(r+q)+r q\left(s_{1}+2\right)+r^{2} s_{2}$ coefficients of the matrices $\mathbf{S}(L), \boldsymbol{\xi}, \boldsymbol{\eta}, \mathbf{B}^{*}$ and $\mathbf{E}(L)$, we assume that $\mathbf{p} \in \Pi$, where $\Pi$ is an open subset of $\mathbb{R}^{\lambda}$ such that for $\mathbf{p} \in \Pi$ : (1) $\mathbf{S}(z)$ is stable, (2) $\operatorname{rank}(\mathbf{B}(z))=q$ with the exception of a finite subset of $\mathbb{C}$, (3) $\operatorname{rank}(\mathbf{B}(1))=\operatorname{rank}\left(\xi \eta^{\prime}\right)=r-c$.

We say that $\mathcal{G}$ is a rational reduced-rank family of filters with cointegrating rank c.
Proposition 3. Assume that $r>q$. Let $\mathbf{y}_{t}$ be a $I(1)$ solution of Equation (10), where $(\mathbf{S}(L), \mathbf{B}(L))$ belongs to a rational reduced-rank family of filters with cointegrating rank c. For generic values of the parameters in $\Pi$, Assumptions 1, 3, 4, 5 and 6 hold. Thus the Strong Form of Proposition 2 holds and $\mathbf{y}_{t}$ has an error correction representation

$$
\mathbf{A}(L) \mathbf{y}_{t}=\mathbf{A}^{*}(L)(1-L) \mathbf{y}_{t}+\boldsymbol{\alpha}\left(\boldsymbol{\beta}^{\prime} \mathbf{y}_{t-1}-\mathbf{w}\right)=\mathbf{B}(0) \mathbf{u}_{t}
$$

where $\mathbf{A}(L)$ is a finite-degree polynomial matrix.

Proof. Part (iii) of Definition 5 implies that Assumptions 1 and 4 hold for all $\mathbf{p} \in \Pi$. The sets where Assumptions 5 and 6 do not hold are the intersections of the open set $\Pi$ with the algebraic varieties

$$
\text { (a) rank }\left[\binom{\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{B}^{*}}{\boldsymbol{\eta}^{\prime}}\right]<q, \quad(\mathrm{~b}) \boldsymbol{\xi}_{\perp}^{\prime}\left(\mathbf{B}^{*}-\mathbf{S}^{*}(1) \mathbf{S}(1)^{-1} \boldsymbol{\xi} \boldsymbol{\eta}^{\prime}\right)=\mathbf{0}
$$

(the variety described by (a) is obtained by equating to zero the determinant of all the $q \times q$ submatrices of the $r \times q$ matrix between brackets). It is easy to see that the varieties (a) and (b) are not trivial,
i.e., that their dimension is lower than $\lambda$. Thus Assumptions 5 and 6 hold generically. The same result holds for Assumption 3. The points of $\Pi$ where it is not fulfilled belong to a lower-dimensional algebraic variety. This is proved in A.1, see in particular Lemma A4.

Remark 7. It is easy to see that, assuming that $c \leq q$, $\operatorname{rank}\left(\boldsymbol{\xi}_{\perp}^{\prime}\left(\mathbf{B}^{*}-\mathbf{S}^{*}(1) \mathbf{S}(1)^{-1} \boldsymbol{\xi} \eta^{\prime}\right)=c\right.$ holds generically in $\Pi$. Thus, in that case, the error term $\beta \mathbf{y}_{t}-\mathbf{w}$ is generically a non-cointegrated $I(0)$ process, see Remark 6.

Remark 8. A general comment on genericity results is in order. Theorems like Proposition 3 or Proposition 1, part (II), show that the subset where some statement does not hold belong to some algebraic variety of lower dimension (see the proof of Proposition 3 in particular), and is therefore negligible from a topological point of view. This suggests the working hypothesis that such subset is negligible from an economic or statistical point of view as well. If, for example, economic theory produces a singular vector $\mathbf{y}_{t}$ with cointegrationg rank $c$, we may find reasonable to conclude that $\mathbf{y}_{t}$ has representation (14) with a finite autoregressive polynomial. However, a greater degree of certainty is obtained by checking that the parameters of $(\mathbf{S}(L), \mathbf{B}(L))$, that are implicit in the theory, do not necessarily lie in one of the three algebraic varieties described in the proof of Proposition 3.

Definition 5 does not assume that $\mathbf{B}(L)$ has no zeros inside the unit circle. Thus we have not assumed that $\mathbf{u}_{t}$ is fundamental for $(1-L) \mathbf{F}_{t}$, see Section 2.2. However, Proposition 3 shows that for generic values of the parameters in $\Pi$, the assumptions of Proposition 2, strong form, hold, Assumption 3 in particular, so that $\mathbf{B}(L)$ has no zeros of non-unit modulus and therefore inside the unit circle. Thus:

Proposition 4. Assume that $r>q$. Let $\mathbf{y}_{t}$ be a solution of Equation (10), where $(\mathbf{S}(L), \mathbf{B}(L))$ belongs to a rational reduced-rank family of filters with cointegrating rank c. For generic values of the parameters in $\Pi, \mathbf{u}_{t}$ is fundamental for $(1-L) \mathbf{y}_{t}$.

Remark 9. Note that Propositions 3 and 4 do not hold in the nonsingular case, where no genericity argument can be used to rule out non-unit zeros of $\mathbf{B}(L)$, either inside or outside the unit circle. In particular, fundamentalness of $\mathbf{u}_{t}$ for $(1-L) \mathbf{y}_{t}$ is not generic if $r=q$.

### 3.3. Permanent and Transitory Shocks

Let $\eta_{\perp}$ be a $q \times d$ matrix whose columns are independent and orthogonal to the columns of $\eta$, and let

$$
\overline{\boldsymbol{\eta}}=\boldsymbol{\eta}\left(\boldsymbol{\eta}^{\prime} \boldsymbol{\eta}\right)^{-1}, \quad \overline{\boldsymbol{\eta}}_{\perp}=\boldsymbol{\eta}_{\perp}\left(\boldsymbol{\eta}_{\perp}^{\prime} \boldsymbol{\eta}_{\perp}\right)^{-1} .
$$

Defining $\mathbf{v}_{1 t}=\eta_{\perp}^{\prime} \mathbf{u}_{t}$, and $\mathbf{v}_{2 t}=\eta^{\prime} \mathbf{u}_{t}$, we have

$$
\mathbf{u}_{t}=\bar{\eta}_{\perp} \mathbf{v}_{1 t}+\bar{\eta} \mathbf{v}_{2 t}=\left(\begin{array}{ll}
\bar{\eta}_{\perp} & \bar{\eta}
\end{array}\right)\binom{\mathbf{v}_{1 t}}{\mathbf{v}_{2 t}}
$$

We have

$$
\mathbf{B}(L) \mathbf{u}_{t}=\left[\mathbf{B}(L)\left(\overline{\boldsymbol{\eta}}_{\perp} \overline{\boldsymbol{\eta}}\right)\right]\binom{\mathbf{v}_{1 t}}{\mathbf{v}_{2 t}}=(1-L) \mathbf{G}_{1}(L) \mathbf{v}_{1 t}+\left(\boldsymbol{\xi}+(1-L) \mathbf{G}_{2}(L)\right) \mathbf{v}_{2 t}
$$

where $\mathbf{G}_{1}(L)=\left(\mathbf{B}^{*}+(1-L) \mathbf{E}(L)\right) \bar{\eta}_{\perp}$, and $\mathbf{G}_{2}(L)=\left(\mathbf{B}^{*}+(1-L) \mathbf{E}(L)\right) \bar{\eta}$. All the solutions of the difference equation $(1-L) \mathbf{y}_{t}=\mathbf{S}(L)^{-1} \mathbf{C}(L) \mathbf{u}_{t}$ are

$$
\begin{equation*}
\mathbf{y}_{t}=\mathbf{S}(L)^{-1}\left[\mathbf{G}_{1}(L) \mathbf{v}_{1 t}+\mathbf{G}_{2}(L) \mathbf{v}_{2 t}+\mathbf{T}_{t}\right]+\mathbf{K}, \tag{21}
\end{equation*}
$$

where $\mathbf{K}$ is a constant stochastic process, and

$$
\mathbf{T}_{t}=\left\{\begin{array}{l}
\boldsymbol{\xi}\left(\mathbf{v}_{21}+\mathbf{v}_{22}+\cdots+\mathbf{v}_{2 t}\right), \text { for } t>0 \\
\mathbf{0}, \text { for } t=0 \\
-\boldsymbol{\xi}\left(\mathbf{v}_{20}+\mathbf{v}_{2,-1}+\cdots+\mathbf{v}_{2, t+1}\right), \text { for } t<0
\end{array}\right.
$$

As $\boldsymbol{\xi}$ is full rank, we see that $\mathbf{y}_{t}$ is driven by the $q-d=r-c$ permanent shocks $\mathbf{v}_{2 t}$, and by the $d$ temporary shocks $\mathbf{v}_{1 t}$. In representation (21), the component $\mathbf{T}_{t}$ is the common-trend of Stock and Watson (1988). Note that the number of permanent shocks is obtained as $r$ minus the cointegrating rank, as usual. However, the number of transitory shocks is only $d=c-(r-q)$, as though $r-q$ transitory shocks had a zero coefficient.

### 3.4. VECMs and Unrestricted VARs in The Levels

Several papers have addressed the issue of whether and when an error correction model or an unrestricted VAR in the levels should be used for estimation in the case of nonsingular cointegrated vectors: Sims et al. (1990) have shown that the parameters of a cointegrated VAR are consistently estimated using an unrestricted VAR in the levels; on the other hand, Phillips (1998) shows that if the variables are cointegrated, the long-run features of the impulse-response functions are consistently estimated only if the unit roots are explicitly taken into account, that is within a VECM specification. The simulation exercise described below provides evidence in favour of the VECM specification in the singular case.
(I) We generate $\mathbf{y}_{t}$ using a specification of (14) with $r=4, q=3, d=2$, so that $c=r-q+d=3$. The $4 \times 4$ matrix $\mathbf{A}(L)$ is of degree 2 . The impulse-response functions are identified by assuming that the upper $3 \times 3$ submatrix of $\mathbf{B}(0)$ is lower triangular (see Appendix $C$ for details). We replicate the generation of $\mathbf{y}_{t} 1000$ times for $T=100,500,1000,5000$.
(II) For each replication, we estimate a (misspecified) VAR in differences (DVAR), a VAR in the levels (LVAR) and a VECM, as in Johansen (1988 1991), assuming known $c$, the degree of $\mathbf{A}(L)$ and that of $\mathbf{A}^{*}(L)$. For the VAR in differences the impulse-response functions for $(1-L) \mathbf{y}_{t}$ are cumulated to obtain impulse-response function for $\mathbf{y}_{t}$. The root mean square error between estimated and actual impulse-response functions is computed for each replication using all 12 impulse-responses and averaged over all replications.

The results are shown in Table 1. We see that the RMSE of both the VECM and the LVAR decreases as $T$ increases. However, for all values of $T$, the RMSE of the VECM stabilizes as the lag increases, whereas it deteriorates for the LVAR, in line with the claim that the long-run response of the variables are better estimated with the VECM. The performance of the misspecified DVAR is uniformly poor with the exception of lag zero.

Table 1. Monte Carlo Simulations. VECM: $r=4, q=3, c=3$.

|  | Lags | DVAR | LVAR | VECM |  | Lags | DVAR | LVAR |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| VECM |  |  |  |  |  |  |  |  |
|  | 0 | 0.06 | 0.05 | 0.05 | 0 | 0.02 | 0.02 | 0.02 |
|  | 4 | 0.26 | 0.18 | 0.17 |  | 4 | 0.23 | 0.07 |
| 100 | 20 | 0.30 | 0.37 | 0.22 | $T=500$ | 20 | 0.25 | 0.14 |
|  | 40 | 0.30 | 0.45 | 0.22 |  | 40 | 0.25 | 0.21 |
|  | 80 | 0.30 | 0.57 | 0.22 |  | 80 | 0.25 | 0.32 |
|  | 0 | 0.02 | 0.02 | 0.02 |  | 0 | 0.01 | 0.01 |
|  | 4 | 0.23 | 0.05 | 0.05 |  | 4 | 0.22 | 0.02 |

Root mean squared errors at different lags, when estimating the impulse-response functions of the simulated variables $\mathbf{y}_{t}$ to the shocks $\mathbf{u}_{t}$. Estimation is carried out using three different autoregressive representations: a VAR for $(1-L) \mathbf{y}_{t}(\mathrm{DVAR})$, a VAR for $\mathbf{y}_{t}(\mathrm{LVAR})$, and a VECM with $c=r-q+d$ error terms (VECM). The results are based on 1000 replications. For the data generating process see Appendix C. The RMSEs are obtained averaging over all replications and all $4 \times 3$ responses.

## 4. Cointegration of the Observable Variables in a DFM

Consider again the factor model $x_{i t}=\chi_{i t}+\epsilon_{i t}$, rewritten here as

$$
\begin{equation*}
\mathbf{x}_{t}=\chi_{t}+\epsilon_{t}, \quad \chi_{t}=\Lambda \mathbf{F}_{t} \tag{22}
\end{equation*}
$$

where $\boldsymbol{\Lambda}$ is $n \times r$, with $n>r$. The relationship between cointegration of the factors $\mathbf{F}_{t}$ and cointegration of the variables $x_{i t}$ is now considered.

Let us recall that the the common factors $\mathbf{F}_{j t}$ are assumed to be orthogonal to the idiosyncratic components $\epsilon_{k s}$ for all $i, j, t, s$, i.e., $\mathrm{E} \chi_{t} \epsilon_{s}^{\prime}=\mathbf{0}_{n \times n}$. for all $t, s$, see the Introduction. The other assumptions on model (22) are asymptotic, see e.g., Forni et al. (2000); Forni and Lippi (2001); (Stock and Watson 2002a, 2002b), and put no restriction on the matrix $\boldsymbol{\Lambda}$ and the vector $\boldsymbol{\epsilon}_{t}$ for a given finite $n$. In particular, the first $r$ eigenvalues of the matrix $\boldsymbol{\Lambda} \boldsymbol{\Lambda}^{\prime}$ must diverge as $n \rightarrow \infty$, but this has no implications on the rank of the matrix $\boldsymbol{\Lambda}$ corresponding to, say, $n=10$. However, as we see in Proposition 5 (iii), if the idiosyncratic components are $I(0)$, then, independently of $\boldsymbol{\Lambda}$, all $p$-dimensional subvectors of $\mathbf{x}_{t}$ are cointegrated for $p>q-d$, which is at odds with what is observed in the macroeconomic datasets analyzed in the empirical Dynamic Factor Model literature. This motivates assuming that $\epsilon_{t}$ is $I(1)$. In that case, see Proposition 5 (i), cointegration of $\mathbf{x}_{t}$ requires that both the common and the idiosyncratic components are cointegrated. Some results are collected in the statement below.

Proposition 5. Let $\mathbf{x}_{t}^{(p)}=\chi_{t}^{(p)}+\boldsymbol{\epsilon}_{t}^{(p)}=\Lambda^{(p)} \mathbf{F}_{t}+\boldsymbol{\epsilon}_{t}^{(p)}$ be a $p$-dimensional subvector of $\mathbf{x}_{t}, p \leq n$. Denote by $c_{\chi}^{p}$ and $c_{\epsilon}^{p}$ the cointegrating rank of $\chi_{t}^{(p)}$ and $\epsilon_{t}^{(p)}$ respectively. Both range from $p$, stationarity, to 0 , no cointegration.
(i) $\mathbf{x}_{t}^{(p)}$ is cointegrated only if $\boldsymbol{\chi}_{t}^{(p)}$ and $\epsilon_{t}^{(p)}$ are both cointegrated.
(ii) If $p>q-d$ then $\chi_{t}^{(p)}$ is cointegrated. If $p \leq q-d$ and $\operatorname{rank}\left(\boldsymbol{\Lambda}^{(p)}\right)<p$ then $\boldsymbol{\chi}_{t}^{(p)}$ is cointegrated.
(iii) Let $V^{\chi} \subseteq \mathbb{R}^{p}$ and $V^{\epsilon} \subseteq \mathbb{R}^{p}$ be the cointegrating spaces of $\chi_{t}^{(p)}$ and $\epsilon_{t}^{(p)}$ respectively. The vector $\mathbf{x}_{t}^{(p)}$ is cointegrated if and only if the intersection of $V^{\chi}$ and $V^{\epsilon}$ contains non-zero vectors. In particular, (a) if $p>q-d$ and $c^{\varepsilon}>q-d$ then $\mathbf{x}^{(p)}$ is cointegrated, (b) if $p>q-d$ and $\boldsymbol{\epsilon}_{t}^{(p)}$ is stationary then $\mathbf{x}^{(p)}$ is cointegrated.

Proof. Because $\chi_{i t}$ and $\epsilon_{j s}$ are orthogonal for all $i, j, t, s$, the spectral densities of $(1-L) \mathbf{x}_{t}^{(p)},(1-L) \boldsymbol{\chi}_{t}^{(p)}$, $(1-L) \boldsymbol{\epsilon}_{t}^{(p)}$ fulfill:

$$
\begin{equation*}
\boldsymbol{\Sigma}_{\Delta x}^{(p)}(\theta)=\boldsymbol{\Sigma}_{\Delta \chi}^{(p)}(\theta)+\boldsymbol{\Sigma}_{\Delta \epsilon}^{(p)}(\theta) \quad \theta \in[-\pi, \pi] . \tag{23}
\end{equation*}
$$

Now, (23) implies that

$$
\begin{equation*}
\lambda_{p}\left(\Sigma_{\Delta x}^{(p)}(0)\right) \geq \lambda_{p}\left(\Sigma_{\Delta \chi}^{(p)}(0)\right)+\lambda^{(p)}\left(\Sigma_{\Delta \epsilon}^{(p)}(0)\right) \tag{24}
\end{equation*}
$$

where $\lambda_{p}(A)$ denotes the smallest eigenvalue of the hermitian matrix $A$; this is one of the Weyl's inequalities, see Franklin (2000), p. 157, Theorem 1. Because the spectral density matrices are non-negative definite, the right hand side in (24) vanishes if and only if both terms on the right hand side vanish, i.e., the spectral density of $\Delta \mathbf{x}_{t}^{(p)}$ is singular at zero if and only if the spectral densities of $\Delta \boldsymbol{\chi}_{t}^{(p)}$ and $\Delta \boldsymbol{\epsilon}_{t}^{(p)}$ are singular at zero. By definition 4, (i) is proved.

Without loss of generality we can assume that $\mathbf{S}(L)=\mathbf{I}_{r}$. By substituting (21) in (22), we obtain

$$
\begin{equation*}
\mathbf{x}_{t}=\boldsymbol{\Lambda}\left[\left(\mathbf{G}_{1}(L) \mathbf{v}_{1 t}+\mathbf{G}_{2}(L) \mathbf{v}_{2 t}+\mathbf{T}_{t}\right)+\mathbf{K}\right]+\boldsymbol{\epsilon}_{t} \tag{25}
\end{equation*}
$$

where on the right hand side the only non-stationary terms are $\mathrm{T}_{t}$ and possibly $\boldsymbol{\epsilon}_{t}$. By recalling that $\mathbf{T}_{t}=\xi \sum_{s=1}^{t} \mathbf{v}_{2 s}$ where $\xi$ is of dimension $r \times(q-d)$ and rank $q-d$, and by defining $\mathcal{G}_{t}=$ $\boldsymbol{\Lambda}\left[\mathbf{G}_{1}(L) \mathbf{v}_{1 t}+\mathbf{G}_{2}(L) \mathbf{v}_{2 t}+\mathbf{K}\right]$ and $\mathcal{T}_{t}=\sum_{s=1}^{t} \mathbf{v}_{2 s}$, we can rewrite (25) as

$$
\mathbf{x}_{t}=\boldsymbol{\Lambda} \xi \mathcal{T}_{t}+\mathcal{G}_{t}+\boldsymbol{\epsilon}_{t}
$$

For $\mathbf{x}_{t}^{(p)}$ :

$$
\mathbf{x}_{t}^{(p)}=\boldsymbol{\chi}_{t}^{(p)}+\boldsymbol{\epsilon}_{t}^{(p)}=\boldsymbol{\Lambda}^{(p)} \boldsymbol{\xi} \mathcal{T}_{t}+\boldsymbol{\mathcal { G }}_{t}^{(p)}+\boldsymbol{\epsilon}_{t}^{(p)}
$$

where $\boldsymbol{\Lambda}^{(p)}$ and $\mathcal{G}_{t}^{(p)}$ have an obvious definition. Of course cointegration of the common components $\boldsymbol{\chi}_{t}^{(p)}$ is equivalent to cointegration of $\boldsymbol{\Lambda}^{(p)} \boldsymbol{\xi} \mathcal{T}_{t}$, which in turn is equivalent to $\operatorname{rank}\left(\boldsymbol{\Lambda}^{(p)} \boldsymbol{\mathcal { F }}\right)<p$. Statement (ii) follows from

$$
\operatorname{rank}\left(\boldsymbol{\Lambda}^{(p)} \boldsymbol{\mathcal { F }}\right) \leq \min \left(\operatorname{rank}\left(\boldsymbol{\Lambda}^{(p)}\right), \operatorname{rank}(\boldsymbol{\mathcal { \xi }})\right)
$$

The first part of (iii) is obvious. Assume now that $p>q-d$. If $c_{\chi}^{p}+c_{\epsilon}^{p}=\operatorname{dim}\left(V^{\chi}\right)+\operatorname{dim}\left(V^{\epsilon}\right)=$ $p-(q-d)+c_{\epsilon}^{p}>p$, i.e., if $c_{\epsilon}^{p}>q-d$, then the intersection between $V^{\chi}$ and $V^{\epsilon}$ is non-trivial, so that $\mathbf{x}_{t}^{(p)}$ is cointegrated.

## 5. Summary and Conclusions

The paper studies representation theory for singular $I(1)$ stochastic vectors, the factors of an $I(1)$ Dynamic Factor Model in particular. Singular $I(1)$ vectors are cointegrated, with a cointegrating rank $c$ equal to $r-q$, the dimension of $\mathbf{y}_{t}$ minus its rank, plus $d$, with $0 \leq d<q$.

If $(1-L) \mathbf{y}_{t}$ has rational spectral density, under assumptions that generalize to the singular case those in Johansen (1995), we show that $\mathbf{y}_{t}$ has an error correction representation with $c$ error terms, thus generalizing the Granger representation theorem (from MA to AR) to the singular case. Important consequences of singularity are that generically: (i) the autoregressive matrix polynomial of the error correction representation is of finite degree, (ii) the white noise vector driving $(1-L) \mathbf{y}_{t}$ is fundamental.

We find that $\mathbf{y}_{t}$ is driven by $r-c$ permanent shocks and $d=c-(r-q)$ transitory shocks, not $c$ as in the nonsingular case.

Using simulated data generated by a simple singular VECM, confirms previous results, obtained for nonsingular vectors, showing that under cointegration the long-run features of impulse-response functions are better estimated using a VECM rather than a VAR in the levels.

In Section 4 we argue that stationarity of the idiosyncratic components in a DFM produce an amount of cointegration for the observable variables $x_{i t}$ that is not observed in the datasets that are standard in applied Dynamic Factor Model literature. Thus the idiosyncratic vector in those datasets is likely to be $I(1)$, so that an estimation strategy robust to the assumption that some of the idiosyncratic variables $\epsilon_{i t}$ are $I(1)$ should be preferred.

The results in this paper are the basis for estimation of $I(1)$ Dynamic Factor Models with cointegrated factors, which is developed in the companion paper (Barigozzi et al. 2019).

Author Contributions: All authors contributed equally to the paper. All authors have read and agreed to the published version of the manuscript.
Funding: This research received no external funding.
Acknowledgments: Dietmar Bauer, Manfred Deistler, Massimo Franchi, Martin Wagner, three anonymous referees and the Editors of this Special Issue gave important suggestions for improvements. We also thank the participants to the Workshop on Estimation and Inference Theory for Cointegrated Processes in the State Space Representation, Technische Universität Dortmund, January 2016. Part of this paper was written while Matteo Luciani was chargé de recherches F.R.S.- F.N.R.S., and he gratefully acknowledges their financial support. Of course we are responsible for any remaining errors.

Disclaimer: The views expressed in this paper are those of the authors and do not necessarily reflect those of the Board of Governors or the Federal Reserve System.

Conflicts of Interest: The authors declare no conflict of interest.

## Appendix A. Proofs

## Appendix A.1. Assumption 3 Holds Generically

Proving that Assumption 3 holds generically is equivalent to proving that $\mathbf{M}(z)$ is generically zeroless, see the argument below Equation (20).

We need some preliminary results. Lemma A1, though quite easy, is not completely standard and is therefore carefully stated and proved below. Regarding notation, to avoid possible misunderstandings, let us recall that vectors and matrices are always denoted by boldface symbols, while light symbols denote scalars, see Lemmas A1 and A2 in particular.

Lemma A1. Let $A_{j}, j=1, \ldots, s$, be scalar polynomials defined on $\mathbb{R}^{\lambda}$, let $\mathbf{p} \in \mathbb{R}^{\lambda}$ and $Q(\mathbf{p})$ be the statement

$$
A_{j}(\mathbf{p})=0, \quad \text { for } j=1, \ldots, s
$$

for example the statement that all the $q \times q$ minors of $\mathbf{M}(1)$ vanish, i.e., that $\operatorname{rank}(\mathbf{M}(1))<q$. Let $\Pi$ be an open subset of $\mathbb{R}^{\lambda}$. If $Q$ is false for one point $\mathbf{p}^{*} \in \mathbb{R}^{\lambda}$, then $Q$ is generically false in $\Pi$.

Proof. Let $\mathcal{N}$ be the closure in $\Pi$ (in the topology of $\Pi$ ) of the subset of $\Pi$ where $Q$ is true. Suppose that $Q$ is not generically false in $\Pi$. Then the interior of $\mathcal{N}$ in $\Pi$, call it $\mathcal{N}^{\circ}$, is not empty. As $\Pi$ is open, $\mathcal{N}^{\circ}$ is open both in the topology of $\Pi$ and of $\mathbb{R}^{\lambda}$. On the other hand a polynomial function defined on $\mathbb{R}^{\lambda}$ vanishes on an open set if and only if it vanishes on the whole $\mathbb{R}^{\lambda}$, which contradicts the existence of a point in $\mathbb{R}^{\lambda}$ where $Q$ is false.

Lemma A2. Consider the scalar polynomials

$$
A(z)=a_{0} z^{n}+a_{1} z^{n-1}+\cdots+a_{n}, \quad B(z)=b_{0} z^{m}+b_{1} z^{m-1}+\cdots+a_{m},
$$

with $a_{0} \neq 0$ and $b_{0} \neq 0$, and let $\alpha_{i}, i=1, \ldots, n$ and $\beta_{j}, j=1, \ldots, m$, be the roots of $A$ and $B$, respectively. Then: (i)

$$
a_{0}^{m} b_{0}^{n} \prod_{i, j}\left(\alpha_{i}-\beta_{j}\right)=R\left(a_{0}, a_{1}, \ldots, a_{n} ; b_{0}, b_{1}, \ldots, b_{m}\right)
$$

where $R$ is a polynomial function which is called the resultant of $A$ and $B$. (ii) The resultant vanishes if and only if $A$ and $B$ have a common root. (iii) Suppose that the coefficients $a_{i}$ and $b_{j}$ are polynomial functions of $\mathbf{p} \in \Pi$, where $\Pi$ is an open subset of $\mathbb{R}^{\lambda}$. If there exists a point $\mathbf{p}^{*} \in \mathbb{R}^{\lambda}$ such that $a_{0}\left(\mathbf{p}^{*}\right) \neq 0, b_{0}\left(\mathbf{p}^{*}\right) \neq 0$, and $R\left(\mathbf{p}^{*}\right) \neq 0$, then generically in $\Pi$ the polynomials $A$ and $B$ have no common roots.

Proof. For (i) and (ii) see van der Waerden (1953, pp. 83-8). Statement (iii) is an obvious consequence of (ii) and Lemma A1.

Lemma A3. Recall that a zero of $\mathbf{M}(z)$ is a complex number $z^{*}$ such that $\operatorname{rank}\left(\mathbf{M}\left(z^{*}\right)\right)<q$. If $\mathbf{M}(z)$ has two $q \times q$ submatrices whose determinants have no common roots, then $\mathbf{M}(z)$ is zeroless.

Proof. If $z^{*}$ is a zero of $\mathbf{M}(z)$, then $z^{*}$ is a zero of all the $q \times q$ submatrices of $\mathbf{M}(z)$.
For the statement and proof of our last result it is convenient to make explicit the dependence of the matrix $\mathbf{M}(z)$ and its submatrices on the vector $\mathbf{p}$. Thus we use $\mathbf{M}^{\mathbf{p}}(z)$, etc. The parameters of the matrix $\mathbf{S}(L)$ play no role here. Hence, with no loss of generality, we assume $s_{2}=0$, so that $\lambda=(r-c)(r+q)+r q\left(s_{1}+2\right)$. Lemmas A2-A4 below imply that Assumption 3 holds generically in $\Pi$.

Lemma A4. Let $\mathbf{M}_{1}^{\mathbf{p}}(z), \mathbf{M}_{2}^{\mathbf{p}}(z), \ldots$ be all the $q \times q$ submatrices of $\mathbf{M}^{\mathbf{p}}(z)$ and let $\mathcal{L}_{i}^{\mathbf{p}}$ be the leading coefficient of $\operatorname{det} \mathbf{M}_{i}^{\mathbf{p}}(z)$ and $R_{i j}^{\mathbf{p}}$ is the resultant of $\operatorname{det} \mathbf{M}_{i}^{\mathbf{p}}(z)$ and $\operatorname{det} \mathbf{M}_{j}^{\mathbf{p}}(z)$. There exist $i, j, \mathbf{p}^{*} \in \mathbb{R}^{\lambda}$ such that

$$
\mathcal{L}_{i}^{\mathrm{p}^{*}} \mathcal{L}_{j}^{\mathrm{p}^{*}} \neq 0
$$

and

$$
R_{i j}^{\mathrm{p}^{*}} \neq 0
$$

Proof. Assume that $r=q+1$. To each $\mathbf{p} \in \Pi$ there corresponds the matrix

$$
\mathbf{M}^{\mathbf{p}}(z)=\binom{\xi_{\perp}^{\prime} \mathbf{B}^{*}}{\boldsymbol{\xi}^{\prime} \boldsymbol{\xi}^{\prime} \eta^{\prime}}+(1-z)\binom{\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{E}(z)}{\xi^{\prime} \mathbf{B}^{*}}+(1-z)^{2}\binom{\mathbf{0}_{c \times q}}{\xi^{\prime} \mathbf{E}(z)} .
$$

Of course, the definition of $\mathbf{M}^{\mathbf{p}}(z)$ makes sense for all $\mathbf{p} \in \mathbb{R}^{\lambda}$, see Equation (19). Let $\mathbf{M}_{1}^{\mathbf{p}}(z)$ and $\mathbf{M}_{2}^{\mathbf{p}}(z)$ be the matrices obtained from $\mathbf{M}^{\mathbf{p}}(z)$ by removing the first and the last row respectively. We have:

$$
\begin{aligned}
& \operatorname{degree}\left[\operatorname{det}\left(\mathbf{M}_{1}^{\mathrm{p}}(z)\right)\right] \leq(q-d)\left(s_{1}+2\right)+d\left(s_{1}+1\right)=d_{1} \\
& \operatorname{degree}\left[\operatorname{det}\left(\mathbf{M}_{2}^{\mathrm{p}}(z)\right)\right] \leq(q-d-1)\left(s_{1}+2\right)+(d+1)\left(s_{1}+1\right)=d_{2}
\end{aligned}
$$

We will construct a point $\mathbf{p}^{*} \in \mathbb{R}^{\lambda}$ such that: (A) the coefficient of $z^{d_{1}}$ in $\operatorname{det}\left(\mathbf{M}_{1}^{\mathbf{p}^{*}}(z)\right)$ and the coefficient of $z^{d_{2}}$ in $\operatorname{det}\left(\mathbf{M}_{2}^{\mathbf{p}^{*}}(z)\right)$ (the leading coefficients) do not vanish, (B) the resultant of $\operatorname{det}\left(\mathbf{M}_{1}^{\mathbf{p}^{*}}(z)\right)$ and $\operatorname{det}\left(\mathbf{M}_{2}^{\mathbf{p}^{*}}(z)\right)$ does not vanish.

Let us firstly define a family of matrices, denoted by $\underline{\mathbf{M}}(z)$, obtained by specifying $\eta, \xi, \xi_{\perp}^{\prime}, \mathbf{B}^{*}$ and $\mathbf{E}(z)$ in the following way:

$$
\begin{aligned}
& \underline{\eta}^{\prime}=\left(\begin{array}{ll}
\mathbf{0}_{(q-d) \times d} & \mathbf{I}_{q-d}
\end{array}\right), \underline{\xi}=\binom{\mathbf{I}_{q-d}}{\mathbf{0}_{c \times(q-d)}}, \underline{\xi}_{\perp}^{\prime}=\binom{\mathbf{K}}{\mathbf{H}}, \\
& \underline{\mathbf{B}}^{*}=\left(\begin{array}{ll}
\mathbf{H}^{\prime} & \mathbf{0}_{(q+1) \times(q-d)}
\end{array}\right), \underline{\mathbf{E}}(z)=\left(\begin{array}{l}
\mathbf{E}_{1}(z) \\
\mathbf{E}_{2}(z) \\
\mathbf{E}_{3}(z)
\end{array}\right),
\end{aligned}
$$

where:

$$
\begin{aligned}
& \mathbf{K}=\left(\begin{array}{lll}
\mathbf{0}_{1 \times(q-d)} & 1 & \mathbf{0}_{1 \times d}
\end{array}\right), \quad \mathbf{H}=\left(\begin{array}{ll}
\mathbf{0}_{d \times(q+1-d)} & \mathbf{I}_{d}
\end{array}\right), \\
& \mathbf{E}_{1}(z)=\left(\begin{array}{ccccc} 
& & & & \\
& k_{1}(z) & h_{1}(z) & \cdots & 0 \\
\mathbf{0}_{(q-d) \times d} & & \ddots & \ddots & \\
& & & \ddots & h_{q-d-1}(z) \\
& 0 & & \cdots & \\
& & & & k_{q-d}(z)
\end{array}\right), \quad \mathbf{E}_{2}(z)=\left(\begin{array}{ll}
e(z) & \left.\mathbf{0}_{1 \times(q-1)}\right), ~
\end{array}\right. \\
& \mathbf{E}_{3}(z)=\left(\begin{array}{ccccc}
f_{1}(z) & g_{1}(z) & \cdots & 0 & \\
& \ddots & \ddots & & \mathbf{0}_{d \times(q-d-1)} \\
0 & \cdots & f_{d}(z) & g_{d}(z) &
\end{array}\right),
\end{aligned}
$$

the entries $e, k_{i}, h_{i}, f_{i}$ and $g_{i}$ being scalar polynomials of degree $s_{1}$.
We denote by $\mathbf{q}_{1}$ the vector including the coefficients of the polynomials $f_{i}, i=1, \ldots, d$ and $k_{i}$, $i=1, \ldots,(q-d)$, a total of $q\left(s_{1}+1\right)$ coefficients, by $\mathbf{q}_{2}$ the vector including the coefficients of the polynomials $e, g_{i}, i=1, \ldots, d$ and $h_{i}, i=1, \ldots,(q-d-1)$, a total of $q\left(s_{1}+1\right)$ coefficients, by $\mathbf{q}_{0}$ the vector including the zeros and the ones in the definition of $\underline{\xi}, \underline{\eta}, \underline{\mathbf{B}}, \underline{\mathbf{E}}$, and define $\mathbf{q}=\left(\boldsymbol{q}_{0} \mathbf{q}_{1} \mathbf{q}_{2}\right)$, which is a $\lambda$-dimensional parameter vector. We put no restriction on $\mathbf{q}_{1}$ and $\mathbf{q}_{2}$, so that both can take any value in $\mathbb{R}^{v}$, with $v=q\left(s_{1}+1\right)$. Note that $\mathbf{q}$ does not necessarily belong to $\Pi$. We have:

$$
\underline{\mathbf{M}}^{\mathbf{q}}(z)=\left(\begin{array}{cc}
\mathbf{0}_{1 \times d} & \mathbf{0}_{1 \times(q-d)}  \tag{A1}\\
\mathbf{I}_{d} & \mathbf{0}_{d \times(q-d)} \\
\mathbf{0}_{(q-d) \times d} & \mathbf{I}_{q-d}
\end{array}\right)+(1-z)\left(\begin{array}{c}
\mathbf{E}_{2}(z) \\
\mathbf{E}_{3}(z) \\
\mathbf{0}_{(q-d) \times q}
\end{array}\right)+(1-z)^{2}\left(\begin{array}{c}
\mathbf{0}_{1 \times q} \\
\mathbf{0}_{d \times q} \\
\mathbf{E}_{1}(z)
\end{array}\right) .
$$

The matrix $\underline{\mathbf{M}}^{\mathbf{q}}(z)$ has zero entries except for the diagonal joining the positions $(1,1)$ and $(q, q)$, and the diagonal joining $(2,1)$ and $(q+1, q)$. The matrices $\underline{\mathbf{M}}_{1}^{\mathbf{q}}(z)$ and $\underline{\mathbf{M}}_{2}^{\mathbf{q}}(z)$ are upper- and lower-triangular, respectively, and

$$
\begin{aligned}
\operatorname{det}\left(\underline{\mathbf{M}}_{1}^{\mathbf{q}}(z)\right)= & {\left[1+(1-z) f_{1}(z)\right] \cdots\left[\left(1+(1-z) f_{d}(z)\right]\right.} \\
& \times\left[1+(1-z)^{2} k_{1}(z)\right] \cdots\left[1+(1-z)^{2} k_{q-d}(z)\right]=\mathcal{L}_{1, d_{1}}^{\mathbf{q}} z^{d_{1}}+\cdots+\mathcal{L}_{1,0}^{\mathbf{q}} \\
\operatorname{det}\left(\underline{\mathbf{M}}_{2}^{\mathbf{q}}(z)\right)= & (1-z)^{2 q-d-1} e(z)\left[g_{1}(z) \cdots g_{d}(z)\right]\left[h_{1}(z) \cdots h_{q-d-1}(z)\right]=\mathcal{L}_{2, d_{2}}^{\mathbf{q}} z^{d_{1}}+\cdots+\mathcal{L}_{2,0}^{\mathbf{q}} .
\end{aligned}
$$

Note that $\operatorname{det}\left(\underline{\mathbf{M}}_{1}^{\mathbf{q}}(z)\right)$ does not depend on $\mathbf{q}_{2}$, while $\operatorname{det}\left(\underline{\mathbf{M}}_{2}^{\mathbf{q}}(z)\right)$ does not depend on $\mathbf{q}_{1}$. Thus we use the notation $\delta_{1}^{\mathbf{q}_{1}}(z)=\operatorname{det}\left(\underline{\mathbf{M}}_{1}^{\mathbf{q}}(z)\right), \delta_{2}^{\mathbf{q}_{2}}(z)=\operatorname{det}\left(\underline{\mathbf{M}}_{2}^{\mathbf{q}}(z)\right), \mathcal{M}_{1, d_{1}}^{\mathbf{q}_{1}}=\mathcal{L}_{1, d_{1}}^{\mathbf{q}}, \mathcal{M}_{2, d_{2}}^{\mathbf{q}_{2}}=\mathcal{L}_{2, d_{2}}^{\mathbf{q}}$. Now:
(i) Let $\mathbf{q}_{2}^{*} \in \mathbf{R}^{v}$ be such that none of the leading coefficients of the polynomials $e, g_{i}$ and $h_{i}$ vanishes. Of course $\mathcal{M}_{2, d_{2}}^{\mathbf{q}_{2}^{*}}=d_{2} \neq 0$.
(ii) Let $\check{z}$ be a root of $\delta_{2}^{\mathbf{q}_{2}^{*}}(z)$. If $\check{z}=1$ then $\check{z}$ is not a root of $\delta_{1}^{\mathbf{q}_{1}}(z)$ for all $\mathbf{q}_{1} \in \mathbb{R}^{v}$. Suppose that $\check{z}$ is a root of $g_{j}(z)$, for some $j$. As the parameters of the polynomials $f_{i}$ and $k_{i}$ are free to vary in $\mathbb{R}^{v}$, then, generically in $\mathbb{R}^{v}, \delta_{1}^{\mathbf{q}_{1}}(\check{z}) \neq 0$. Iterating for all roots of $\delta_{2}^{\mathbf{q}_{2}^{*}}(z)$, generically in $\mathbb{R}^{v}, \delta_{1}^{\mathbf{q}_{1}}(z)$ and $\delta_{2}^{\mathbf{q}_{2}^{*}}(z)$ have no roots in common. Moreover, generically in $\mathbb{R}^{v}, \mathcal{M}_{1, d_{1}}^{\boldsymbol{q}_{1}}=d_{1} \neq 0$. Thus, there exists $\mathbf{q}_{1}^{*}$ such that (a) $\mathcal{M}_{1, d_{1}}^{\mathbf{q}_{1}^{*}}=d_{1} \neq 0$, (b) $\delta_{1}^{\mathbf{q}_{1}^{*}}(z)$ and $\delta_{2}^{\mathbf{q}_{2}^{*}}(z)$ have no roots in common.
(iii) Now let $\mathbf{p}^{*}=\left(\mathbf{q}_{0} \mathbf{q}_{1}^{*} \mathbf{q}_{2}^{*}\right)$, so that

$$
\left.\operatorname{det}\left(\mathbf{M}_{1}^{\mathrm{p}^{*}}(z)\right)=\delta_{1}^{\mathbf{q}_{1}^{*}}(z)\right), \quad \operatorname{det}\left(\mathbf{M}_{2}^{\mathrm{p}^{*}}(z)\right)=\delta_{2}^{\mathbf{q}_{2}^{*}}(z)
$$

Using (i) and (ii), (A) the leading coefficients of $\operatorname{det}\left(\mathbf{M}_{1}^{\mathrm{p}^{*}}(z)\right)$ and $\operatorname{det}\left(\mathbf{M}_{2}^{\mathrm{p}^{*}}(z)\right)$ do not vanish, (B) $\operatorname{det}\left(\mathbf{M}_{1}^{\mathbf{p}^{*}}(z)\right)$ and $\operatorname{det}\left(\mathbf{M}_{2}^{\mathbf{p}^{*}}(z)\right)$ have no root in common so that their resultant does not vanish. This proves the proposition for $r=q+1$.

Generalizing this result to $r>q+1$ is easy. Let us define the family $\underline{\mathbf{N}}(z)$ in the following way: (a) specify $\eta^{\prime}, \boldsymbol{\xi}, \mathbf{E}_{1}(z)$ and $\mathbf{E}_{3}(z)$ as in the definition of $\underline{\mathbf{M}}(z)$, (b) then let

$$
\left.\begin{array}{rl}
\mathbf{K} & =\left(\begin{array}{lll}
\mathbf{0}_{(r-q) \times(r-d-1)} & \left(\begin{array}{l}
\mathbf{0}_{1 \times(r-q-1)} 1
\end{array}\right)^{\prime} \quad \mathbf{0}_{(r-q) \times d}
\end{array}\right), \quad \mathbf{H}=\left(\begin{array}{ll}
\mathbf{0}_{d \times(r-d)} & \mathbf{I}_{d}
\end{array}\right), \\
\underline{\underline{\xi}}_{\perp}^{\prime} & =\binom{\mathbf{K}}{\mathbf{H}}, \underline{\mathbf{D}}=\left(\begin{array}{ll}
\mathbf{H}^{\prime} & \mathbf{I}_{r \times(q-d)}
\end{array}\right), \quad \mathbf{E}_{2}(z)=\left(\begin{array}{c}
\mathbf{0}_{(r-q) \times q} \\
(e(z) \\
\mathbf{0}_{1 \times(q-1)}
\end{array}\right)
\end{array}\right) .
$$

We have:

$$
\underline{\mathbf{N}}(z)=\left(\begin{array}{cc}
\mathbf{0}_{(r-q) \times d} & \mathbf{0}_{(r-q) \times(q-d)} \\
\mathbf{I}_{d} & \mathbf{0}_{d \times(q-d)} \\
\mathbf{0}_{(q-d) \times d} & \mathbf{I}_{q-d}
\end{array}\right)+(1-z)\left(\begin{array}{c}
\mathbf{E}_{2}(z) \\
\mathbf{E}_{3}(z) \\
\mathbf{0}_{(q-d) \times q}
\end{array}\right)+(1-z)^{2}\left(\begin{array}{c}
\mathbf{0}_{(r-q) \times q} \\
\mathbf{0}_{d \times q} \\
\mathbf{E}_{1}(z)
\end{array}\right)
$$

It is easy to see that the $(q+1) \times q$ lower submatrix of $\underline{\mathbf{N}}(z)$ is identical to the matrix $\underline{\mathbf{M}}^{\mathbf{q}}(z)$ in (A1).
Appendix A.2. if $R>Q$ and $C \leq Q$, Assumptions 5 and 6 Do Not Imply That $\mathbf{e}_{t}$ Is a Non-Cointegrated $I(0)$ Process.

Let $r=3, q=2, \mathbf{S}(L)=\mathbf{I}_{3}$,

$$
\xi=\left(\begin{array}{l}
1 \\
0 \\
0
\end{array}\right), \quad \eta=\binom{0}{1}, \quad \xi_{\perp}=\left(\begin{array}{ll}
0 & 0 \\
1 & 0 \\
0 & 1
\end{array}\right), \quad \mathbf{B}^{*}=\left(\begin{array}{ll}
a & b \\
1 & 0 \\
1 & 0
\end{array}\right)
$$

In this case $c=2$ and $d=1$, so that $c=q$ (see Remark 6). We have

$$
\binom{\xi_{\perp}^{\prime} \mathbf{B}^{*}}{\eta^{\prime}}=\left(\begin{array}{cc}
1 & 0 \\
1 & 0 \\
\hline 0 & 1
\end{array}\right)
$$

We see that Assumptions 5 and 6 hold. However, $\operatorname{rank}\left(\boldsymbol{\xi}_{\perp}^{\prime} \mathbf{B}^{*}\right)=1$, so that $\mathbf{e}_{t}$, though being $I(0)$, is not a non-cointegrated $I(0)$ process. On the other hand, if the $(3,2)$ entry of $\mathbf{B}^{*}$ is 1 instead of 0 , $\mathbf{e}_{t}$ is non-cointegrated.

## Appendix B. Non Uniqueness

In Proposition 3 we prove that a singular $I(1)$ vector with cointegrating rank $c$ has a finite error correction representation with $c$ error terms. On the other hand, as we have seen in Remark 5, when $c=r-q$ the singular vector $\mathbf{y}_{t}$ has also an autoregressive representation in the differences, i.e., a representation with zero error terms. In Appendix B. 1 we give an example hinting that $\mathbf{y}_{t}$ has error correction representations with any number of error terms between $d$ and $c$. However, in Appendix B. 2 we show that all such representations produce the same impulse-response functions.

## Appendix B.1. Alternative Representations with Different Numbers of Error Terms

Let $\mathbf{S}(L)=\mathbf{I}_{r}$ and consider the following example, with $r=3, q=2, c=2$, so that $d=1$ :

$$
\begin{aligned}
\xi^{\prime} & =\left(\begin{array}{lll}
1 & 1 & 1
\end{array}\right) \\
\eta^{\prime} & =\left(\begin{array}{llr}
1 & 2
\end{array}\right) \\
\xi_{\perp}^{\prime} & =\left(\begin{array}{lrr}
1 & -1 & 0 \\
0 & 1 & -1
\end{array}\right)
\end{aligned}
$$

We have,

$$
(1-L)\binom{\xi_{\perp}^{\prime}}{\hat{\zeta}^{\prime}} \mathbf{y}_{t}=\left(\begin{array}{ccc}
1-L & 0 & 0 \\
0 & 1-L & 0 \\
0 & 0 & 1
\end{array}\right)\left\{\left(\begin{array}{cc}
b_{11}^{*}-b_{21}^{*} & b_{12}^{*}-b_{22}^{*} \\
b_{21}^{*}-b_{31}^{*} & b_{22}^{*}-b_{32}^{*} \\
3 & 6
\end{array}\right)+(1-L) \hat{\mathbf{E}}(L)\right\} \mathbf{u}_{t},
$$

where $(1-L) \hat{\mathbf{E}}(L)$ gathers the second and third terms in $\mathbf{M}(L)$. If the assumptions of Proposition 2 hold, we obtain an error correction representation with error terms

$$
\xi_{\perp}^{\prime} \mathbf{y}_{t}=\binom{y_{1 t}-y_{2 t}}{y_{2 t}-y_{3 t}} .
$$

However, we also have

$$
\begin{aligned}
& (1-L)\binom{\xi_{1}^{\prime}}{z^{\prime}} \mathbf{y}_{t}=\left(\begin{array}{ccc}
1-L & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right) \\
& \times\left\{\left(\begin{array}{cc}
b_{11}^{*}-b_{21}^{*} & b_{12}^{*}-b_{22}^{*} \\
(1-L)\left(b_{21}^{*}-b_{31}^{*}\right) & (1-L)\left(b_{22}^{*}-b_{32}^{*}\right) \\
3 & 6
\end{array}\right)+(1-L) \check{\mathbf{E}}(L)\right\} \mathbf{u}_{t}=\left(\begin{array}{ccc}
1-L & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right) \check{\mathbf{M}}(L) \mathbf{u}_{t} .
\end{aligned}
$$

Under suitable assumptions on the coefficients $b_{i j}^{*}$ and $\check{\mathbf{E}}(L)$, assuming in particular that the matrix

$$
\left(\begin{array}{cc}
b_{11}^{*}-b_{21}^{*} & b_{12}^{*}-b_{22}^{*} \\
3 & 6
\end{array}\right)
$$

is nonsingular, the matrix $\check{\mathbf{M}}(L)$ is zeroless and has therefore a finite-degree left inverse. Proceeding as in Proposition 2, we obtain an alternative error correction representation with just one error term, namely $y_{1 t}-y_{2 t}$.

This example should be sufficient to convey the idea that $y_{t}$ admits error correction representations with a minimum $d$ and a maximum $c=r-q+d$ of error terms.

The problem of error correction representations, with different numbers of error terms, has been recently addressed in Deistler and Wagner (2017). An implication of their main result (see Theorem 1, p. 41) is that if $\mathbf{y}_{t}$ has the error correction representation

$$
\tilde{\mathbf{A}}(L) \mathbf{y}_{t}=\tilde{\mathbf{A}}^{*}(L)(1-L) \mathbf{y}_{t}+\tilde{\mathbf{A}}(1) \mathbf{y}_{t-1}=\tilde{\mathbf{B}} \tilde{\mathbf{u}}_{t},
$$

and $\operatorname{rank}(\tilde{\mathbf{A}}(1))<c$ (the number of error terms is not the maximum), then $\tilde{\mathbf{A}}(L)$ and $\tilde{\mathbf{B}}$ are not left coprime.

The consequences of Deistler and Wagner's paper have not yet been developed. In Propositions 2 and 3 we have only considered representations with $c$ error terms. On non-uniqueness of autoregressive representations for singular vectors with rational spectral density see also Chen et al. (2011); Anderson et al. (2012); Forni et al. (2015).

## Appendix B.2. Uniqueness of Impulse-Response Functions

Suppose that the assumptions of Proposition 2, weak form, hold. Let $\mathbf{y}_{t}$ be a solution of Equation (10), so that

$$
\begin{equation*}
(1-L) \mathbf{y}_{t}=\mathbf{S}(L)^{-1} \mathbf{B}(L) \mathbf{u}_{t} \tag{A2}
\end{equation*}
$$

and suppose that $\mathbf{y}_{t}$ has the autoregressive representation

$$
\begin{equation*}
\tilde{\mathbf{A}}(L) \mathbf{y}_{t}=\tilde{\mathbf{B}} \tilde{\mathbf{u}}_{t} \tag{A3}
\end{equation*}
$$

where $\tilde{\mathbf{A}}(L)$ is a rational matrix with poles outside the unit circle, $\tilde{\mathbf{A}}(0)=\mathbf{I}_{r}, \tilde{\mathbf{u}}_{t}$ is a nonsingular $q$-dimensional white noise, $\tilde{\mathbf{B}}$ is a full rank $r \times q$ matrix $^{5}$. We have

$$
\begin{equation*}
\tilde{\mathbf{A}}(L)\left[(1-L) \mathbf{y}_{t}\right]=(1-L) \tilde{\mathbf{B}} \tilde{\mathbf{u}}_{t} . \tag{A4}
\end{equation*}
$$

The assumption that $\tilde{\mathbf{B}}$ is full rank and the argument used e.g., in Brockwell and Davis (1991), p. 111, Problem 3.8, imply that $\tilde{\mathbf{u}}_{t}$ is fundamental for $(1-L) \mathbf{y}_{t}$. Thus $\tilde{\mathbf{u}}_{t}=\mathbf{Q} \mathbf{u}_{t}$, where $\mathbf{Q}$ is a nonsingular $q \times q$ matrix (see Rozanov (1967), p. 57), and $\tilde{\mathbf{B}} \tilde{\mathbf{u}}_{t}=[\tilde{\mathbf{B}} \mathbf{Q}] \mathbf{u}_{t}$.

On the other hand, from (A2) and (A4):

$$
\begin{equation*}
\tilde{\mathbf{A}}(L) \mathbf{S}(L)^{-1} \mathbf{B}(L) \mathbf{u}_{t}=(1-L)[\tilde{\mathbf{B}} \mathbf{Q}] \mathbf{u}_{t} . \tag{A5}
\end{equation*}
$$

As $\mathbf{u}_{t}$ is nonsingular, $\tilde{\mathbf{A}}(L) \mathbf{S}(L)^{-1} \mathbf{B}(L)=(1-L)[\tilde{\mathbf{B}} \mathbf{Q}]$. Setting $L=0$ we have $\tilde{\mathbf{B}} \mathbf{Q}=\mathbf{B}(0)$, so that (A3) becomes

$$
\begin{equation*}
\tilde{\mathbf{A}}(L) \mathbf{y}_{t}=\mathbf{B}(0) \mathbf{u}_{t} \tag{A6}
\end{equation*}
$$

while (A5) becomes

$$
\begin{equation*}
\tilde{\mathbf{A}}(L) \mathbf{S}(L)^{-1} \mathbf{B}(L) \mathbf{u}_{t}=(1-L) \mathbf{B}(0) \mathbf{u}_{t} . \tag{A7}
\end{equation*}
$$

The impulse-response function of $\mathbf{y}_{t}$ to $\mathbf{u}_{t}$ resulting from (A6) is $\mathbf{H}(L) \mathbf{B}(0)$, where $\mathbf{H}(L) \tilde{\mathbf{A}}(L)=\mathbf{I}_{r}$. Multiplying both sides of (A7) by $\mathbf{H}(L)$ we obtain

$$
\mathbf{S}(L)^{-1} \mathbf{B}(L)=(1-L) \mathbf{H}(L) \mathbf{B}(0)
$$

so that $\mathbf{H}(L) \mathbf{B}(0)$ is obtained by cumulating $\mathbf{S}(L)^{-1} \mathbf{B}(L)$ and is therefore independent of $\tilde{\mathbf{A}}(L)$.

## Appendix C. Data Generating Process for the Simulations

The simulation results of Section 3.4 are obtained using the following specification of (14):

$$
\mathbf{A}(L) \mathbf{y}_{t}=\mathbf{A}^{*}(L)(1-L) \mathbf{y}_{t}+\boldsymbol{\alpha} \boldsymbol{\beta}^{\prime} \mathbf{y}_{t-1}=\mathbf{C}(0) \mathbf{u}_{t}=\mathbf{G H} \mathbf{u}_{t},
$$

where $r=4, q=3, c=3$, the degree of $\mathbf{A}(L)$ is 2 , so that the degree of $\mathbf{A}^{*}(L)$ is $1 . \mathbf{A}(L)$ is generated using the factorization

$$
\mathbf{A}(L)=\boldsymbol{\mathcal { U }}(L) \boldsymbol{\mathcal { M }}(L) \boldsymbol{\mathcal { V }}(L),
$$

where $\mathcal{U}(L)$ and $\mathcal{V}(L)$ are $r \times r$ matrix polynomials with all their roots outside the unit circle, and

$$
\mathcal{M}(L)=\left(\begin{array}{cc}
(1-L) \mathbf{I}_{r-c} & \mathbf{0} \\
\mathbf{0} & \mathbf{I}_{c}
\end{array}\right)
$$

[^14](see Watson 1994). To get a $\operatorname{VAR}(2)$ we set $\mathcal{U}(L)=\mathbf{I}_{r}-\mathcal{U}_{1} L$, and $\mathcal{V}(L)=\mathbf{I}_{r}$, and then, by rewriting $\mathcal{M}(L)=\mathbf{I}_{r}-\mathcal{M}_{1} L$, we get $\mathbf{A}_{1}=\mathcal{M}_{1}+\mathcal{U}_{1}$, and $\mathbf{A}_{2}=-\mathcal{M}_{1} \mathcal{U}_{1}$.

Regarding the generation of the data, the diagonal entries of the matrix $\mathcal{U}_{1}$ are drawn from a uniform distribution between 0.5 and 0.8 , while the extra-diagonal entries are drawn from a uniform distribution between 0 and 0.3. $\mathcal{U}_{1}$ is then multiplied by a scalar so that its largest eigenvalue is 0.6 . The matrix $\mathbf{G}$ is generated as in Bai and Ng (2007): (1) $\tilde{\mathbf{G}}$ is an $r \times r$ diagonal matrix of rank $q$ where $\tilde{g}_{i i}$ is drawn from the uniform distribution between 0.8 and $1.2,(2) \mathrm{G}$ is obtained by orthogonalizing an $r \times r$ uniform random matrix, (3) $\mathbf{G}$ is equal to the first $q$ columns of the matrix $\tilde{G}^{\mathbf{G}} \tilde{G}^{1 / 2}$. Lastly, the orthogonal matrix $\mathbf{H}$ is such that the upper $3 \times 3$ submatrix of $\mathbf{G H}$ is lower triangular. The results are based on 1000 replications. The matrices $\mathcal{U}_{1}, \mathrm{G}$ and $\mathbf{H}$ are generated only once (the numerical values are available on request) so that the set of impulse responses to be estimated is the same for all replications, whereas the vector $\mathbf{u}_{t}$ is redrawn from $\mathcal{N}\left(\mathbf{0}, \mathbf{I}_{4}\right)$ at each replication.

## References

Amengual, Dante, and Mark W. Watson. 2007. Consistent estimation of the number of dynamic factors in a large $N$ and $T$ panel. Journal of Business and Economic Statistics 25: 91-96.
Anderson, Brian DO, and Manfred Deistler. 2008a. Generalized linear dynamic factor models-A structure theory. Paper presented at IEEE Conference on Decision and Control, Cancun, Mexico, December 9-11.
Anderson, Brian DO, and Manfred Deistler. 2008b. Properties of zero-free transfer function matrices. SICE Journal of Control, Measurement and System Integration 1: 284-92.
Anderson, Brian DO, Manfred Deistler, Weitian Chen, and Alexander Filler. 2012. Autoregressive models of singular spectral matrices. Automatica 48: 2843-49.
Bai, Jushan, and Serena Ng. 2007. Determining the number of primitive shocks in factor models. Journal of Business and Economic Statistics 25: 52-60.
Banerjee, Anindya, Massimiliano Marcellino, and Igor Masten. 2014. Forecasting with factor-augmented error correction models. International Journal of Forecasting 30: 589-612.
Banerjee, Anindya, Massimiliano Marcellino, and Igor Masten. 2017. Structural FECM: Cointegration in large-scale structural FAVAR models. Journal of Applied Econometrics 32: 1069-86.
Barigozzi, Matteo, Antonio M. Conti, and Matteo Luciani. 2014. Do euro area countries respond asymmetrically to the common monetary policy? Oxford Bulletin of Economics and Statistics 76: 693-714.
Barigozzi, Matteo, Marco Lippi, and Matteo Luciani. 2019. Large-dimensional dynamic factor models: Estimation of impulse-response functions with $I(1)$ cointegrated factors. arXiv arXiv:1602:02398.
Bauer, Dietmar, and Martin Wagner. 2012. A State Space Canonical Form For Unit Root Processes. Econometric Theory 28: 1313-49.
Brockwell, Peter J., and Richard A. Davis. 1991. Time Series: Theory and Methods, 2nd ed. New York: Springer.
Canova, Fabio. 2007. Methods for Applied Macroeconomics. Princeton: Princeton University Press.
Chen, Weitian, Brian DO Anderson, Manfred Deistler, and Alexander Filler. 2011. Solutions of Yule-Walker equations for singular AR processes. Journal of Time Series Analysis 32: 531-38.
Deistler, Manfred, Brian DO Anderson, A. Filler, Ch. Zinner, and W. Chen. 2010. Generalized linear dynamic factor models: An approach via singular autoregressions. European Journal of Control 16: 211-24.
Deistler, Manfred, and Martin Wagner. 2017. Cointegration in singular ARMA models. Economics Letters 155: 39-42.
Forni, Mario, and Luca Gambetti. 2010. The dynamic effects of monetary policy: A structural factor model approach. Journal of Monetary Economics 57: 203-16.
Forni, Mario, Domenico Giannone, Marco Lippi, and Lucrezia Reichlin. 2009. Opening the Black Box: Structural Factor Models versus Structural VARs. Econometric Theory 25: 1319-47.
Forni, Mario, Marc Hallin, Marco Lippi, and Lucrezia Reichlin. 2000. The Generalized Dynamic Factor Model: Identification and Estimation. The Review of Economics and Statistics 82: 540-54.
Forni, Mario, Marc Hallin, Marco Lippi, and Paolo Zaffaroni. 2015. Dynamic factor models with infinite-dimensional factor spaces: One-sided representations. Journal of Econometrics 185: 359-71.
Forni, Mario, and Marco Lippi. 2001. The Generalized Dynamic Factor Model: Representation Theory. Econometric Theory 17: 1113-41.

Franchi, Massimo, and Paolo Paruolo. 2019. A general inversion theorem for cointegration. Econometric Reviews 38: 1176-201.
Franklin, J. N. 2000. Matrix Theory, 2nd ed. New York: Dover Publications.
Giannone, Domenico, Lucrezia Reichlin, and Luca Sala. 2005. Monetary policy in real time. In NBER Macroeconomics Annual 2004. Edited by Mark Gertler and Kenneth Rogoff. Cambridge: MIT Press, chp. 3, pp. 161-224.
Gregoir, Stéphane. 1999. Multivariate Time Series With Various Hidden Unit Roots, Part I. Econometric Theory 15: 435-68.
Johansen, Søren. 1988. Statistical analysis of cointegration vectors. Journal of Economic Dynamics and Control 12: 231-54.
Johansen, Søren. 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica 59: 1551-80.
Johansen, Søren. 1995. Likelihood-Based Inference in Cointegrated Vector Autoregressive Models, 1st ed. Oxford: Oxford University Press.
Lancaster, Peter, and Miron Tismenetsky. 1985. The Theory of Matrices, 2nd ed. New York: Academic Press.
Luciani, Matteo. 2015. Monetary policy and the housing market: A structural factor analysis. Journal of Applied Econometrics 30: 199-218.
Phillips, Peter C.B. 1998. Impulse response and forecast error variance asymptotics in nonstationary VARs. Journal of Econometrics 83: 21-56.
Rozanov, Yu. A. 1967. Stationary Random Processes. San Francisco: Holden-Day.
Sargent, Thomas J. 1989. Two Models of Measurements and the Investment Accelerator. Journal of Political Economy 97: 251-87.
Sims, Christopher, James H. Stock, and Mark W. Watson. 1990. Inference in linear time series models with some unit roots. Econometrica 58: 113-44.
Stock, James H., and Mark W. Watson. 1988. Testing for common trends. Journal of the American Statistical Association 83: 1097-107.
Stock, James H., and Mark W. Watson. 2002a. Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association 97: 1167-79.
Stock, James H., and Mark W. Watson. 2002b. Macroeconomic forecasting using diffusion indexes. Journal of Business and Economic Statistics 20: 147-62.
Stock, James H., and Mark W. Watson. 2016. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In Handbook of Macroeconomics. Edited by John B. Taylor and Harald Uhlig. Amsterdam: North Holland, Elsevier, vol. 2A, chp. 8, pp. 415-525.
Van der Waerden, Bartel Leendert. 1953.Modern Algebra, 2nd ed. New York: Frederick Ungar, vol. I.
Watson, Mark W. 1994. Vector autoregressions and cointegration. In Handbook of Econometrics. Edited by Robert F. Engle and Daniel L. McFadden. Amsterdam: North Holland, Elsevier, vol. 4, chp. 47, pp. 2843-915.
© 2020 by Matteo Barigozzi and Marco Lippi. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http:/ / creativecommons.org/licenses/by/4.0/).

## Article

# A Parameterization of Models for Unit Root Processes: Structure Theory and Hypothesis Testing 

Dietmar Bauer ${ }^{1, *}$, Lukas Matuschek ${ }^{1,2}$, Patrick de Matos Ribeiro ${ }^{1,2}$ and Martin Wagner 3,4,5<br>1 Faculty of Business Administration and Economics, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany; matuschek@statistik.tu-dortmund.de (L.M.); ribeiro@statistik.tu-dortmund.de (P.d.M.R.)<br>2 Faculty of Statistics, TU Dortmund, Vogelpothsweg 87, 44227 Dortmund, Germany<br>3 Department of Economics, University of Klagenfurt, Universitätsstraße 65-67, 9020 Klagenfurt, Austria; Martin.Wagner@aau.at<br>4 Bank of Slovenia, Slovenska 35, 1505 Ljubljana, Slovenia<br>5 Institute for Advanced Studies, Josefstädter Straße 39, 1080 Vienna, Austria<br>* Correspondence: Dietmar.Bauer@uni-bielefeld.de; Tel.: +49-521-106-4873

Received: 18 April 2018; Accepted: 4 November 2020; Published: 10 November 2020


#### Abstract

We develop and discuss a parameterization of vector autoregressive moving average processes with arbitrary unit roots and (co)integration orders. The detailed analysis of the topological properties of the parameterization-based on the state space canonical form of Bauer and Wagner (2012)-is an essential input for establishing statistical and numerical properties of pseudo maximum likelihood estimators as well as, e.g., pseudo likelihood ratio tests based on them. The general results are exemplified in detail for the empirically most relevant cases, the (multiple frequency or seasonal) $\mathrm{I}(1)$ and the I(2) case. For these two cases we also discuss the modeling of deterministic components in detail.


Keywords: canonical form; cointegration; hypothesis testing; parameterization; state space representation; unit roots

## 1. Introduction

Since the seminal contribution of Clive W.J. Granger (1981) that introduced the concept of cointegration, the modeling of multivariate (economic) time series with models and methods that allow for unit roots and cointegration has become standard econometric practice with applications ranging from macroeconomics to finance to climate science.

The most prominent (parametric) model class for cointegration analysis are vector autoregressive (VAR) models, popularized by the important contributions of Søren Johansen and Katarina Juselius and their co-authors, see, e.g., the monographs Johansen (1995) and Juselius (2006). The popularity of VAR cointegration analysis stems not only from the (relative) simplicity of the model class, but also from the fact that the VAR cointegration literature is very well-developed and provides a large battery of tools for diagnostic testing, impulse response analysis, forecast error variance decompositions and the like. All this makes VAR cointegration analysis to a certain extent the benchmark in the literature. ${ }^{1}$

The imposition of specific cointegration properties on an estimated VAR model becomes increasingly complicated as one moves away from the I(1) case. As discussed in Section 2, e.g., in the

[^15]I(2) case a triple of indices needs to be chosen (fixed or determined via testing) to describe the cointegration properties. The imposition of cointegration properties in the estimation algorithm then leads to "switching" type algorithms that come together with non-trivial parameterization restrictions involving non-linear inter-relations, compare Paruolo (1996) or Paruolo (2000). ${ }^{2}$ Mathematically, these complications arise from the fact that the unit root and cointegration properties are in the VAR setting related to rank restrictions on the autoregressive polynomial matrix and its derivatives.

Restricting cointegration analysis to VAR processes may be too restrictive. First, it is well-known since Zellner and Palm (1974) that VAR processes are not invariant with respect to marginalization, i.e., subsets of the variables of a VAR process are in general vector autoregressive moving average (VARMA) processes. Second, similar to the first argument, aggregation of VAR processes also leads to VARMA processes, an issue relevant, e.g., in the context of temporal aggregation and in mixed-frequency settings. Third, the linearized solutions to dynamic stochastic general equilibrium (DSGE) models are typically VARMA rather than VAR processes, see, e.g., Campbell (1994). Fourth, a VARMA model may be a more parsimonious description of the data generating process (DGP) than a VAR model, with parsimony becoming more important with increasing dimension of the process. ${ }^{3}$

If one accepts the above arguments as a motivation for considering VARMA processes in cointegration analysis, it is convenient to move to the-essentially equivalent (see Hannan and Deistler 1988, chps. 1 and 2)—state space framework. A key challenge when moving from VAR to VARMA models-or state space models-is that identification becomes an important issue for the latter model class, whereas unrestricted VAR models are (reduced-form) identified. In other words, there are so-called equivalence classes of VARMA models that lead to the same dynamic behavior of the observed process. As is well-known, to achieve identification, restrictions have to be placed on the coefficient matrices in the VARMA case, e.g., zero or exclusion restrictions. A mapping attaching to every transfer function, i.e, the function relating the error sequence to the observed process, a unique VARMA (or state space) system from the corresponding class of observationally equivalent systems is called canonical form. Since not all entries of the coefficient matrices in canonical form are free parameters, for statistical analysis a so-called parameterization is required that maps the free parameters from coefficient matrices in canonical form into a parameter vector. These issues, including the importance of the properties such as continuity and differentiability of parameterizations, are discussed in detail in Hannan and Deistler (1988, chp. 2) and, of course, are also relevant for our setting in this paper.

The convenience of the state space framework for unit root and cointegration analysis stems from the fact that (static and dynamic) cointegration can be characterized by orthogonality constraints, see Bauer and Wagner (2012), once an appropriate basis for the state vector, which is a (potentially singular) VAR process of order one, is chosen. The integration properties are governed by the eigenvalue structure of unit modulus eigenvalues of the system matrix in the state equation. Eigenvalues of unit modulus and orthogonality constraints arguably are easier restrictions to deal with or to implement than the interrelated rank restrictions considered in the VAR or VARMA setting. The canonical form of Bauer and Wagner (2012) is designed for cointegration analysis by using a basis of the state vector that puts the unit root and cointegration properties to the center and forefront. Consequently, these results are key input for the present paper and are thus briefly reviewed in Section 3.

[^16]An important problem with respect to appropriately defining the "free parameters" in VARMA models is the fact that no continuous parameterization of all VARMA or state space models of a certain order $n$ exists in the multivariate case (see Hazewinkel and Kalman 1976). This implies that the model set, $M_{n}$ say, has to be partitioned into subsets on which continuous parameterizations exist, i.e., $M_{n}=\bigcup_{\Gamma \in G} M_{\Gamma}$ for some multi-index $\Gamma$ varying in an index set $G$. Based on the canonical form of Bauer and Wagner (2012), the partitioning is according to systems-in addition to other restrictions such as fixed order $n$-with fixed unit root properties, to be precise over systems with given state space unit root structure. This has the advantage that, e.g., pseudo maximum likelihood (PML) estimation can straightforwardly be performed over systems with fixed unit root properties without any further ado, i.e., without having to consider (or ignore) rank restrictions on polynomial matrices. The definition and detailed discussion of the properties of this parameterization is the first main result of the paper.

The second main set of results, provided in Section 4, is a detailed discussion of the relationships between the different subsets of models $M_{\Gamma}$ for different indices $\Gamma$ and the parameterization of the respective model sets. Knowledge concerning these relations is important to understand the asymptotic behavior of PML estimators and pseudo likelihood ratio tests based on them. In particular, the structure of the closures of $M, \bar{M}$ say, of the considered model set $M$ has to be understood, since the difference $\bar{M} \backslash M$ cannot be avoided when maximizing the pseudo likelihood function ${ }^{4}$. Additionally, the inclusion properties between different sets $M_{\Gamma}$ need to be understood, as this knowledge is important for developing hypothesis tests, in particular for developing hypothesis tests for the dimensions of cointegrating spaces. Hypotheses testing, with a focus on the MFI(1) and I(2) cases, is discussed in Section 5, which shows how the parameterization results of the paper can be used to formulate a large number of hypotheses on (static and polynomial) cointegrating relationships as considered in the VAR cointegration literature. This discussion also includes commonly used deterministic components such as intercept, seasonal dummies, and linear trend, as well as restrictions on these components.

The paper is organized as follows: Section 2 briefly reviews VAR and VARMA models with unit roots and cointegration and discusses some of the complications arising in the VARMA case in addition to the complications arising due to the presence of unit roots and cointegration already in the VAR case. Section 3 presents the canonical form and the parameterization based on it, with the discussion starting with the multiple frequency $\mathrm{I}(1)$ - $\mathrm{MFI}(1)$-and $\mathrm{I}(2)$ cases prior to a discussion of the general case. This section also provides several important definitions like, e.g., of the state space unit root structure. Section 4 contains a detailed discussion concerning the topological structure of the model sets and Section 5 discusses testing of a large number of hypotheses on the cointegrating spaces commonly tested in the cointegration literature. The discussion in Section 5 focuses on the empirically most relevant $\operatorname{MFI}(1)$ and $\mathrm{I}(2)$ cases and includes the usual deterministic components considered in the literature. Section 6 briefly summarizes and concludes the paper. All proofs are relegated to the Appendices $A$ and $B$.

Throughout we use the following notation: $L$ denotes the lag operator, i.e., $L\left(\left\{x_{t}\right\}_{t \in \mathbb{Z}}\right):=$ $\left\{x_{t-1}\right\}_{t \in \mathbb{Z}}$, for brevity written as $L x_{t}=x_{t-1}$. For a matrix $\gamma \in \mathbb{C}^{s \times r}, \gamma^{\prime} \in \mathbb{C}^{r \times s}$ denotes its conjugate transpose. For $\gamma \in \mathbb{C}^{s \times r}$ with full column rank $r<s$, we define $\gamma_{\perp} \in \mathbb{C}^{s \times(s-r)}$ of full column rank such that $\gamma^{\prime} \gamma_{\perp}=0$. $I_{p}$ denotes the $p$-dimensional identity matrix, $0_{m \times n}$ the $m$ times $n$ zero matrix. For two matrices $A \in \mathbb{C}^{m \times n}, B \in \mathbb{C}^{k \times l}, A \otimes B \in \mathbb{C}^{m k \times n l}$ denotes the Kronecker product of $A$ and $B$. For a complex valued quantity $x, \mathcal{R}(x)$ denotes its real part, $\mathcal{I}(x)$ its imaginary part and $\bar{x}$ its complex conjugate. For a set $V, \bar{V}$ denotes its closure. ${ }^{5}$ For two sets $V$ and $W, V \backslash W$ denotes the difference of $V$ and $W$, i.e., $\{v \in V: v \notin W\}$. For a square matrix $A$ we denote the spectral radius (i.e., the maximum of the moduli of its eigenvalues) by $\lambda_{|\max |}(A)$ and by $\operatorname{det}(A)$ its determinant.

[^17]
## 2. Vector Autoregressive, Vector Autoregressive Moving Average Processes and Parameterizations

In this paper, we define VAR processes $\left\{y_{t}\right\}_{t \in \mathbb{Z}}, y_{t} \in \mathbb{R}^{s}$, as solution of

$$
\begin{equation*}
a(L) y_{t}=y_{t}+\sum_{j=1}^{p} a_{j} y_{t-j}=\varepsilon_{t}+\Phi d_{t} \tag{1}
\end{equation*}
$$

with $a(L):=I_{s}+\sum_{j=1}^{p} a_{j} L^{j}$, where $a_{j} \in \mathbb{R}^{s \times s}$ for $j=1, \ldots, p, \Phi \in \mathbb{R}^{s \times m}, a_{p} \neq 0$, a white noise process $\left\{\varepsilon_{t}\right\}_{t \in \mathbb{Z}}, \varepsilon_{t} \in \mathbb{R}^{s}$, with $\Sigma:=\mathbb{E}\left(\varepsilon_{t} \varepsilon_{t}^{\prime}\right)>0$ and a vector sequence $\left\{d_{t}\right\}_{t \in \mathbb{Z}}, d_{t} \in \mathbb{R}^{m}$, comprising deterministic components like, e.g., the intercept, seasonal dummies or a linear trend. Furthermore, we impose the non-explosiveness condition $\operatorname{det} a(z) \neq 0$ for all $|z|<1$, with $a(z):=$ $I_{s}+\sum_{j=1}^{p} a_{j} z^{j}$ and $z$ denoting a complex variable. ${ }^{6}$

Thus, for given autoregressive order $p$, with—as defining characteristic of the order— $a_{p} \neq 0$, the considered class of VAR models with specified deterministic components $\left\{d_{t}\right\}_{t \in \mathbb{Z}}$ is given by the set of all polynomial matrices $a(z)$ such that (i) the non-explosiveness condition holds, (ii) $a(0)=I_{s}$ and (iii) $a_{p} \neq 0$; together with the set of all matrices $\Phi \in \mathbb{R}^{s \times m}$.

Equivalently, the model class can be characterized by a set of rational matrix functions $k(z):=a(z)^{-1}$, referred to as transfer functions, and the input-output description for the deterministic variables, i.e.,

$$
\begin{aligned}
V_{p, \Phi} & :=V_{p} \times \mathbb{R}^{s \times m} \\
V_{p} & :=\left\{k(z)=\sum_{j=0}^{\infty} k_{j} z^{j}=a(z)^{-1}: a(z)=I_{s}+\sum_{j=1}^{p} a_{j} z^{j}, \operatorname{det} a(z) \neq 0 \text { for }|z|<1, a_{p} \neq 0\right\}
\end{aligned}
$$

The associated parameter space is $\Theta_{p, \Phi}:=\Theta_{p} \times \mathbb{R}^{s m} \subset \mathbb{R}^{s^{2} p+s m}$, where the parameters

$$
\begin{equation*}
\boldsymbol{\theta}:=\left[\boldsymbol{\theta}_{a}^{\prime}, \boldsymbol{\theta}_{\Phi}^{\prime}\right]^{\prime}=\left[\operatorname{vec}\left(a_{1}\right)^{\prime}, \ldots, \operatorname{vec}\left(a_{p}\right)^{\prime}, \operatorname{vec}(\Phi)^{\prime}\right]^{\prime} \tag{2}
\end{equation*}
$$

are obtained from stacking the entries of the matrices $a_{j}$ and $\Phi$, respectively.
Remark 1. In the above discussion the parameters, $\theta_{\Sigma}$ say, describing the variance covariance matrix $\Sigma$ of $\varepsilon_{t}$ are not considered. These can be easily included, similarly to $\Phi$ by, e.g., parameterizing positive definite symmetric $s \times s$ matrices via their lower triangular Cholesky factor. This leads to a parameter space $\Theta_{p, \Phi, \Sigma} \subset$ $\mathbb{R}^{s^{2} p+s m+\frac{s(s+1)}{2}}$. We omit $\theta_{\Sigma}$ for brevity, since typically no cross-parameter restrictions involving parameters corresponding to $\Sigma$ are considered, whereas as discussed in Section 5 parameter restrictions involving-in this paper in the state space rather than the VAR setting-both elements of $\Theta_{p}$ and $\Phi$, to, e.g., impose the absence of a linear trend in the cointegrating space, are commonly considered in the cointegration literature. ${ }^{7}$ The estimator of the variance covariance matrix $\Sigma$ often equals the sample variance of suitable residuals $\hat{\varepsilon}_{t}(\theta)$ from (1), if there are no cross-restrictions between $\theta$ and $\theta_{\Sigma}$. This holds, e.g., for the Gaussian pseudo maximum likelihood estimator. Thus, explicitly including $\theta_{\Sigma}$ and $\Theta_{\Sigma}$ in the discussion would only overload notation without adding any additional insights, given the simple nature of the parameterization of $\Sigma$.

[^18]Remark 2. Our consideration of deterministic components is a special case of including exogenous variables. We include exogenous deterministic variables with a static input-output behavior governed solely by the matrix $\Phi$. More general exogenous variables that are dynamically related to the output $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ could be considered, thereby considering so-called VARX models rather than VAR models, which would necessitate considering in addition to the transfer function $k(z)$ also a transfer function $l(z)$, say, linking the exogenous variables dynamically to the output.

For the VAR case, the fact that the mapping assigning a given transfer function $k(z) \in V_{p}$, to a parameter vector $\boldsymbol{\theta}_{a} \in \Theta_{p}$-the parameterization-is continuous with continuously differentiable inverse is immediate. ${ }^{8}$ Homeomorphicity of a parameterization is important for the properties of parameter estimators, e.g., the ordinary least squares (OLS) or Gaussian PML estimator, compare the discussion in Hannan and Deistler (1988, Theorem 2.5.3 and Remark 1, p. 65).

For OLS estimation one typically considers the larger set $V_{p}^{O L S}$ without the non-explosiveness condition and without the assumption $a_{p} \neq 0$ :

$$
V_{p}^{\text {OLS }}:=\left\{k(z)=\sum_{j=0}^{\infty} k_{j} z^{j}=a(z)^{-1}: a(z)=I_{s}+\sum_{j=1}^{p} a_{j} z^{j}\right\} .
$$

Considering $V_{p}^{O L S}$ allows for unconstrained optimization. It is well-known that for $\left\{\varepsilon_{t}\right\}_{t \in \mathbb{Z}}$ as given above, the OLS estimator is consistent over the larger set $V_{p}^{O L S}$, i.e., without imposing non-explosiveness and also when specifying $p$ too high. Alternatively, and closely related to OLS in the VAR case, the pseudo likelihood can be maximized over $\Theta_{p, \Phi}$. With this approach, maxima respectively suprema can occur at the boundary of the parameter space, i.e., maximization effectively has to consider $\bar{\Theta}_{p, \Phi}$. It is well-known that the PML estimator is consistent for the stable case (cf. Hannan and Deistler 1988, Theorem 4.2.1), but the maximization problem is complicated by the restrictions on the parameter space stemming from the non-explosiveness condition. Avoiding these complications and asymptotic equivalence of OLS and PML in the stable VAR case explains why VAR models are usually estimated by OLS. ${ }^{9}$

To be more explicit, ignore deterministic components for a moment and consider the case where the DGP is a stationary VAR process, i.e., a solution of $(1)$ with $a(z)$ satisfying the stability condition $\operatorname{det} a(z) \neq 0$ for $|z| \leq 1$. Define the corresponding set of stable transfer functions by $V_{p, \bullet}$ :

$$
V_{p, \bullet}:=\left\{a(z)^{-1} \in V_{p}: \operatorname{det} a(z) \neq 0 \text { for }|z| \leq 1, a_{p} \neq 0\right\}
$$

Clearly, $V_{p, \bullet}$ is an open subset of $V_{p}$. If the DGP is a stationary VAR process, the above-mentioned consistency result of the OLS estimator over $V_{p}^{O L S}$ implies that the probability that the estimated transfer function, $\hat{k}(z)=\hat{a}(z)^{-1}$ say, is contained in $V_{p, \bullet}$ converges to one as the sample size tends to infinity. Moreover, the asymptotic distribution of the estimated parameters is normal, under appropriate assumptions on $\left\{\varepsilon_{t}\right\}_{t \in \mathbb{Z}}$.

The situation is a bit more involved if the transfer function of the DGP corresponds to a point in the set $\bar{V}_{p, \bullet} \backslash V_{p, \bullet}$, which contains systems with unit roots, i.e., determinantal roots of $a(z)$ on the unit circle, as well as lower order autoregressive systems-with these two cases non-disjoint. The stable lower order case is relatively unproblematic from a statistical perspective. If, e.g., OLS estimation is performed over $V_{p}^{O L S}$, while the true model corresponds to an element in $V_{p^{*}, \bullet}$, with $p^{*}<p$, the OLS estimator is

[^19]still consistent, since $V_{p^{*}, \bullet} \subset V_{p}^{O L S}$. Furthermore, standard chi-squared pseudo likelihood ratio test based inference still applies. The integrated case, for a precise definition see the discussion below Definition 1, is a bit more difficult to deal with, as in this case not all parameters are asymptotically normally distributed and nuisance parameters may be present. Consequently, parameterizations that do not take the specific nature of unit root processes into account are not very useful for inference in the unit root case, see, e.g., Sims et al. (1990, Theorem 1). Studying the unit root and cointegration properties is facilitated by resorting to suitable parameterizations that "zoom in on the relevant characteristics".

In case that the only determinantal root of $a(z)$ on the unit circle is at $z=1$, the system corresponds to a so-called $I(d)$ process, with the integration order $d>0$ made precise in Definition 1 below. Consider first the $\mathrm{I}(1)$ case: As is well-known, the rank of the matrix $a(1)$ equals the dimension of the cointegrating space given in Definition 3 below-also referred to as the cointegrating rank. Therefore, determination of the rank of this matrix is of key importance. With the parameterization used so far, imposing a certain (maximal) rank on $a(1)$ implies complicated restrictions on the matrices $a_{j}, j=1, \ldots, p$. This in turn renders the correspondingly restricted optimization unnecessarily complicated and not conducive to develop tests for the cointegrating rank. It is more convenient to consider the so-called vector error correction model (VECM) representation of autoregressive processes, discussed in full detail in the monograph Johansen (1995). To this end let us first introduce the differencing operator at frequency $0 \leq \omega \leq \pi$

$$
\Delta_{\omega}:= \begin{cases}I_{s}-2 \cos (\omega) L+L^{2} & \text { for } 0<\omega<\pi  \tag{3}\\ I_{s}-\cos (\omega) L & \text { for } \omega \in\{0, \pi\}\end{cases}
$$

For notational brevity, we omit the dependence on $L$ in $\Delta_{\omega}(L)$, henceforth denoted as $\Delta_{\omega}$. Using this notation, the $\mathrm{I}(1)$ error correction representation is given by

$$
\begin{align*}
\Delta_{0} y_{t} & =\Pi y_{t-1}+\sum_{j=1}^{p-1} \Gamma_{j} \Delta_{0} y_{t-j}+\varepsilon_{t}+\Phi d_{t}  \tag{4}\\
& =\alpha \beta^{\prime} y_{t-1}+\sum_{j=1}^{p-1} \Gamma_{j} \Delta_{0} y_{t-j}+\varepsilon_{t}+\Phi d_{t}
\end{align*}
$$

with the matrix $\Pi:=-a(1)=-\left(I_{s}+\sum_{j=1}^{p} a_{j}\right)$ of rank $0 \leq r \leq s$ factorized into the product of two full rank matrices $\alpha, \beta \in \mathbb{R}^{s \times r}$ and $\Gamma_{j}:=\sum_{m=j+1}^{p} a_{m}, j=1, \ldots, p-1$.

This constitutes a reparameterization, where $k(z) \in V_{p}$ is now represented by the matrices $\left(\alpha, \beta, \Gamma_{1}, \ldots, \Gamma_{p-1}\right)$ and a corresponding parameter vector $\boldsymbol{\theta}_{a}^{\mathrm{VECM}} \in \Theta_{p, r}^{\mathrm{VECM}}$. Please note that stacking the entries of the matrices does not lead to a homeomorphic mapping from $V_{p}$ to $\Theta_{p, s}^{\mathrm{VECM}}$, since for $0<r \leq s$ the matrices $\alpha$ and $\beta$ are not identifiable from the product $\alpha \beta^{\prime}$, since $\alpha \beta^{\prime}=\alpha M M^{-1} \beta^{\prime}=\tilde{\alpha} \tilde{\beta}^{\prime}$ for all regular matrices $M \in \mathbb{R}^{r \times r}$. One way to obtain identifiability is to introduce the restriction $\beta=\left[I_{r}, \beta^{* \prime}\right]^{\prime}$, with $\beta^{*} \in \mathbb{R}^{(s-r) \times r}$ and $\alpha \in \mathbb{R}^{s \times r}$. With this additional restriction the parameter vector $\boldsymbol{\theta}_{a}^{\mathrm{VECM}}$ is given by stacking the vectorized matrices $\alpha, \beta^{*}, \Gamma_{1}, \ldots, \Gamma_{p-1}$, similarly to (2). Then $\Theta_{p, r, \Phi}^{\mathrm{VECM}}=\Theta_{p, r}^{\mathrm{VECM}} \times \mathbb{R}^{s m} \subset$ $\mathbb{R}^{p s^{2}-(s-r)^{2}+s m}$. Note for completeness that the normalization of $\beta=\left[I_{r}, \beta^{* \prime}\right]^{\prime}$ may necessitate a re-ordering of the variables in $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ since-without potential reordering-this parameterization implies a restriction of generality as, e.g., processes, where the first variable is integrated, but does not cointegrate with the other variables, cannot be represented.

Define the following sets of transfer functions:

$$
\begin{aligned}
V_{p, r} & :=\left\{a(z)^{-1} \in V_{p}: \operatorname{det} a(z) \neq 0 \text { for }\{z:|z|=1, z \neq 1\}, \operatorname{rank}(a(1)) \leq r\right\}, \\
V_{p, r}^{R R R} & :=\left\{a(z)^{-1} \in V_{p}^{O L S}: \operatorname{rank}(a(1)) \leq r\right\} .
\end{aligned}
$$

The dimension of the parameter vector $\boldsymbol{\theta}_{a}^{\text {VECM }}$ depends on the dimension of the cointegrating space, thus the parameterization of $k(z) \in V_{p, r}$ depends on $r$. The so-called reduced rank regression (RRR) estimator, given by the maximizer of the pseudo likelihood over $V_{p, r}^{R R R}$ is consistent, see, e.g., Johansen (1995, chp. 6). The RRR estimator uses an "implicit" normalization of $\beta$ and thereby implicitly addresses the mentioned identification problem. However, for testing hypotheses involving the free parameters in $\alpha$ or $\beta$, typically the identifying assumption given above is used, as discussed in Johansen (1995, chp. 7).

Furthermore, since $V_{p, r} \subset V_{p, r^{*}}$ for $r<r^{*} \leq s$, with $\Theta_{p, r}^{\mathrm{VECM}}$ a lower dimensional subset of $\Theta_{p, r^{*}}^{\mathrm{VECM}}$, pseudo likelihood ratio testing can be used to sequentially test for the rank $r$, starting with the hypothesis of a rank $r=0$ against the alternative of a rank $0<r \leq s$, and increasing the assumed rank consecutively until the null hypothesis is not rejected.

Ensuring that $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ generated from (4) is indeed an $\mathrm{I}(1)$ process, requires on the one hand that $\Pi$ is of reduced rank, i.e., $r<s$ and on the other that the matrix

$$
\begin{equation*}
\alpha_{\perp}^{\prime} \Gamma \beta_{\perp}:=\alpha_{\perp}^{\prime}\left(I_{s}-\sum_{j=1}^{p-1} \Gamma_{j}\right) \beta_{\perp} \tag{5}
\end{equation*}
$$

has full rank. It is well-known that condition (5) is fulfilled on the complement of a "thin" algebraic subset of $V_{p, r}^{R R R}$, and is therefore, ignored in estimation, as it is "generically" fulfilled. ${ }^{10}$

The $\mathrm{I}(2)$ case is similar in structure to the $\mathrm{I}(1)$ case, but with two rank restrictions and one full rank condition to exclude even higher integration orders. The corresponding VECM is given by

$$
\begin{equation*}
\Delta_{0}^{2} y_{t}=\alpha \beta^{\prime} y_{t-1}-\Gamma \Delta_{0} y_{t-1}+\sum_{j=1}^{p-2} \Psi_{j} \Delta_{0}^{2} y_{t-j}+\varepsilon_{t} \tag{6}
\end{equation*}
$$

with $\alpha, \beta$ as defined in (4), $\Gamma$ as defined in (5) and $\Psi_{j}:=-\sum_{k=j+1}^{p-1} \Gamma_{k}, j=1, \ldots, p-2$. From (5) we already know that reduced rank of

$$
\begin{equation*}
\alpha_{\perp}^{\prime} \Gamma \beta_{\perp}=: \quad \xi \eta^{\prime} \tag{7}
\end{equation*}
$$

with $\xi, \eta \in \mathbb{R}^{(s-r) \times m}, m<s-r$ is required for higher integration orders. The condition for the corresponding solution process $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ to be an $I(2)$ process is given by full rank of

$$
\xi_{\perp}^{\prime} \alpha_{\perp}^{\prime}\left(\Gamma \beta\left(\beta^{\prime} \beta\right)^{-1}\left(\alpha^{\prime} \alpha\right)^{-1} \alpha^{\prime} \Gamma+I_{s}-\sum_{j=1}^{p-2} \Psi_{j}\right) \beta_{\perp} \eta_{\perp}
$$

which again is typically ignored in estimation, just like condition (5) in the $\mathrm{I}(1)$ case. Thus, $\mathrm{I}(2)$ processes correspond to a "thin subset" of $V_{p, r}^{R R R}$, which in turn constitutes a "thin subset" of $V_{p}^{O L S}$. The fact that integrated processes correspond to "thin sets" in $V_{p}^{O L S}$ implies that obtaining estimated systems with specific integration and cointegration properties requires restricted estimation based on parameterizations tailor made to highlight these properties.

Already for the $I(2)$ case, formulating parameterizations that allow conveniently studying the integration and cointegration properties is a quite challenging task. Johansen (1997) contains several different (re-)parameterizations for the $I(2)$ case and Paruolo (1996) defines "integration indices", $r_{0}, r_{1}, r_{2}$ say, as the number of columns of the matrices $\beta \in \mathbb{R}^{s \times r_{0}}, \beta_{1}:=\beta_{\perp} \eta \in \mathbb{R}^{s \times r_{1}}$ and $\beta_{2}:=$ $\beta_{\perp} \eta_{\perp} \in \mathbb{R}^{s \times r_{2}}$. Clearly, the indices $r_{0}, r_{1}, r_{2}$ are linked to the ranks of the above matrices $\Pi$ and $\alpha_{\perp}^{\prime} \Gamma \beta_{\perp}$, as $r_{0}=r$ and $r_{1}=m$ and the columns of $\left[\beta, \beta_{1}, \beta_{2}\right]$ form a basis of $\mathbb{R}^{s}$, such that $s=r_{0}+r_{1}+r_{2}$.

[^20]It holds that $\left\{\beta_{2}^{\prime} y_{t}\right\}_{t \in \mathbb{Z}}$ is an $\mathrm{I}(2)$ process without cointegration and $\left\{\beta_{1}^{\prime} y_{t}\right\}_{t \in \mathbb{Z}}$ is an $\mathrm{I}(1)$ process without cointegration. The process $\left\{\beta^{\prime} y_{t}\right\}_{t \in \mathbb{Z}}$ is typically $I(1)$ and in this case cointegrates with $\left\{\beta_{2}^{\prime} \Delta_{0} y_{t}\right\}_{t \in \mathbb{Z}}$ to stationarity. Thus, there is a direct correspondence of these indices to the dimensions of the different cointegrating spaces-both static and dynamic (with precise definitions given below in Definition 3). ${ }^{11}$ Please note that again, as already before in the $\mathrm{I}(1)$ case, different values of the integration indices $r_{0}, r_{1}, r_{2}$, lead to parameter spaces of different dimensions. Furthermore, in these parameterizations matrices describing different cointegrating spaces are (i) not identified and (ii) linked by restrictions, compare the discussion in Paruolo (2000, sct. 2.2) and (7). These facts render the analysis of the cointegration properties in I(2) VAR systems complicated. Also, in the I(2) VAR case usually some forms of RRR estimators are considered over suitable subsets $V_{p, r, m}^{R R R}$ of $V_{p, r}^{R R R}$, again based on implicit normalizations. Inference, however, again requires one to consider parameterizations explicitly.

Estimation and inference issues are fundamentally more complex in the VARMA case than in the VAR case. This stems from the fact that unrestricted estimation-unlike in the VAR case-is not possible due to a lack of identification, as discussed below. This means that in the VARMA case identification and parameterization issues need to be tackled as the first step, compare the discussion in Hannan and Deistler (1988, chp. 2).

In this paper, we consider VARMA processes as solutions of the vector difference equation

$$
y_{t}+\sum_{j=1}^{p} a_{j} y_{t-j}=\varepsilon_{t}+\sum_{j=1}^{q} b_{j} \varepsilon_{t-j}+\Phi d_{t}
$$

with $a(L):=I_{s}+\sum_{j=1}^{p} a_{j} L^{j}$, where $a_{j} \in \mathbb{R}^{s \times s}$ for $j=1, \ldots, p, a_{p} \neq 0$ and the non-explosiveness condition $\operatorname{det}(a(z)) \neq 0$ for $|z|<1$. Similarly, $b(L):=I_{s}+\sum_{j=1}^{q} b_{j} L^{j}$, where $b_{j} \in \mathbb{R}^{s \times s}$ for $j=1, \ldots, q$, $b_{q} \neq 0$ and $\Phi \in \mathbb{R}^{s \times m}$. The transfer function corresponding to a VARMA process is $k(z):=a(z)^{-1} b(z)$.

It is well-known that without further restrictions the VARMA realization $(a(z), b(z))$ of the transfer function $k(z)=a(z)^{-1} b(z)$ is not identified, i.e., different pairs of polynomial matrices $(a(z), b(z))$ can realize the same transfer function $k(z)$. It is clear that $k(z)=a(z)^{-1} m(z)^{-1} m(z) b(z)=a(z)^{-1} b(z)$ for all non-singular polynomial matrices $m(z)$. Thus, the mapping $\pi$ attaching the transfer function $k(z)=a(z)^{-1} b(z)$ to the pair of polynomial matrices $(a(z), b(z))$ is not injective. ${ }^{12}$

Consequently, we refer for given rational transfer function $k(z)$ to the class $\{(a(z), b(z)): k(z)=$ $\left.a(z)^{-1} b(z)\right\}$ as a class of observationally equivalent VARMA realizations of $k(z)$. To achieve identification requires to define a canonical form, selecting one member of each class of observationally equivalent VARMA realizations for a set of considered transfer functions. A first step towards a canonical form is to only consider left coprime pairs $(a(z), b(z)) \cdot{ }^{13}$ However, left coprimeness is not sufficient for identification and thus further restrictions are required, leading to parameter vectors of smaller dimension than $\mathbb{R}^{s^{2}(p+q)}$. A widely used canonical form is the (reverse) echelon canonical form, see Hannan and Deistler (1988, Theorem 2.5.1, p. 59), based on (monic) normalizations of the diagonal elements of $a(z)$ and degree relationships between diagonal and off-diagonal elements as well as the entries in $b(z)$, which lead to zero restrictions. The (reverse) echelon canonical form in conjunction with a transformation to an error correction model was used in VARMA cointegration analysis in the I(1) case, e.g., in Poskitt (2006, Theorem 4.1), but, as for the VAR case, understanding the interdependencies of rank conditions already becomes complicated once one moves to the $\mathrm{I}(2)$ case.

[^21]In the VARMA case matters are further complicated by another well-known problem that makes statistical analysis considerably more involved compared to the VAR case. Although there exists a generalization of the autoregressive order to the VARMA case, such that any transfer function corresponding to a VARMA system has an $\operatorname{order} n \in \mathbb{N}$ (with the precise definition given in the next section) it is known since Hazewinkel and Kalman (1976) that no continuous parameterization of all rational transfer functions of order $n$ exists if $s>1$. Therefore, if one wants to keep the above-discussed advantages that continuity of a parameterization provides, the set of transfer functions of order $n$, henceforth referred to as $M_{n}$, has to be partitioned into sets on which continuous parameterizations exist, i.e., $M_{n}=\bigcup_{\Gamma \in G} M_{\Gamma}$, for some index set $G$, as already mentioned in the introduction. ${ }^{14}$ For any given partitioning of the set $M_{n}$ it is important to understand the relationships between the different subsets $M_{\Gamma}$, as well as the closures of the pieces $M_{\Gamma}$, since in case of misspecification of $M_{\Gamma}$ points in $\bar{M}_{\Gamma} \backslash M_{\Gamma}$ cannot be avoided even asymptotically in, e.g., pseudo maximum likelihood estimation. These are more complicated issues in the VARMA case than in the VAR case, see the discussion in Hannan and Deistler (1988, Remark 1 after Theorem 2.5.3).

Based on these considerations, the following section provides and discusses a parameterization that focuses on unit root and cointegration properties, resorting to the state space framework that-as mentioned in the introduction-provides advantages for cointegration analysis. In particular, we derive an almost everywhere homeomorphic parameterization, based on partitioning the set of all considered transfer functions according to a multi-index $\Gamma$ that contains, among other elements, the state space unit root structure. This implies that certain cointegration properties are invariant for all systems corresponding to a subset $M_{\Gamma}$, i.e., the parameterization allows to directly impose cointegration properties such as the "cointegration indices" of Paruolo (1996) mentioned before.

## 3. The Canonical Form and the Parameterization

As a first step we define the class of VARMA processes considered in this paper, using the differencing operator defined in (3):

Definition 1. The s-dimensional real VARMA process $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ has unit root structure $\Omega:=$ $\left(\left(\omega_{1}, h_{1}\right), \ldots,\left(\omega_{l}, h_{l}\right)\right)$ with $0 \leq \omega_{1}<\omega_{2}<\cdots<\omega_{l} \leq \pi, h_{k} \in \mathbb{N}, k=1, \ldots, l, l \geq 1$, if it is a solution of the difference equation

$$
\begin{equation*}
\Delta_{\Omega}\left(y_{t}-\Phi d_{t}\right):=\prod_{k=1}^{l} \Delta_{\omega_{k}}^{h_{k}}\left(y_{t}-\Phi d_{t}\right)=v_{t} \tag{8}
\end{equation*}
$$

where $\left\{d_{t}\right\}_{t \in \mathbb{Z}}$ is an m-dimensional deterministic sequence, $\Phi \in \mathbb{R}^{s \times m}$ and $\left\{v_{t}\right\}_{t \in \mathbb{Z}}$ is a linearly regular stationary VARMA process, i.e., there exists a pair of left coprime matrix polynomials $(a(z), b(z)), \operatorname{det} a(z) \neq 0$, $|z| \leq 1$ such that $v_{t}=a(L)^{-1} b(L)\left(\varepsilon_{t}\right)=: c(L)\left(\varepsilon_{t}\right)$ for a white noise process $\left\{\varepsilon_{t}\right\}_{t \in \mathbb{Z}}$ with $\mathbb{E}\left(\varepsilon_{t} \varepsilon_{t}^{\prime}\right)=\Sigma>0$, with furthermore $c(z) \neq 0$ for $z=e^{i \omega_{k}}, k=1, \ldots, l$.

- The process $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ is called unit root process with unit roots $z_{k}:=e^{i \omega_{k}}$ for $k=1, \ldots, l$, the set $F(\Omega):=\left\{\omega_{1}, \ldots, \omega_{l}\right\}$ is the set of unit root frequencies and the integers $h_{k}, k=1, \ldots, l$ are the integration orders.
- A unit root process with unit root structure $((0, d)), d \in \mathbb{N}$, is an $\mathrm{I}(\mathrm{d})$ process.
- A unit root process with unit root structure $\left(\left(\omega_{1}, 1\right), \ldots,\left(\omega_{l}, 1\right)\right)$ is an $\operatorname{MFI}(1)$, process.

A linearly regular stationary VARMA process has empty unit root structure $\Omega_{0}:=\{ \}$.

[^22]As discussed in Bauer and Wagner (2012) the state space framework is convenient for the analysis of VARMA unit root processes. Detailed treatments of the state space framework are given in Hannan and Deistler (1988) and—in the context of unit root processes-Bauer and Wagner (2012).

A state space representation of a unit root VARMA process is ${ }^{15}$

$$
\begin{align*}
y_{t} & =C x_{t}+\Phi d_{t}+\varepsilon_{t} \\
x_{t+1} & =A x_{t}+B \varepsilon_{t} \tag{9}
\end{align*}
$$

for a white noise process $\left\{\varepsilon_{t}\right\}_{t \in \mathbb{Z}}, \varepsilon_{t} \in \mathbb{R}^{s}$, a deterministic process $\left\{d_{t}\right\}_{t \in \mathbb{Z}}, d_{t} \in \mathbb{R}^{m}$ and the unobserved state process $\left\{x_{t}\right\}_{t \in \mathbb{Z}}, x_{t} \in \mathbb{C}^{n}, A \in \mathbb{C}^{n \times n}, B \in \mathbb{C}^{n \times s}, C \in \mathbb{C}^{s \times n}$ and $\Phi \in \mathbb{R}^{s \times m}$.

Remark 3. Bauer and Wagner (2012, Theorem 2) show that every real valued unit root VARMA process $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ as given in (8) has a real valued state space representation with $\left\{x_{t}\right\}_{t \in \mathbb{Z}}$ real valued and real valued system matrices ( $A, B, C$ ). Considering complex valued state space representations in (9) is merely for algebraic convenience, as in general some eigenvalues of $A$ are complex valued. Note for completeness that Bauer and Wagner (2012) contains a detailed discussion why considering the A-matrix in the canonical form in (up to reordering) the Jordan normal form is useful for cointegration analysis. For the sake of brevity we abstain from including this discussion again in the present paper. The key aspect of this construction is its usefulness for cointegration analysis, which becomes visible in Remark 4, where the "simple" unit root properties of blocks of the state vector are discussed.

The transfer function $k(z)$ with real valued power series coefficients corresponding to a real valued unit root process $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ as given in Definition 1 is given by the rational matrix function $k(z)=\Delta_{\Omega}(z)^{-1} a(z)^{-1} b(z)$. The (possibly complex valued) matrix triple ( $A, B, C$ ) realizes the transfer function $k(z)$ if and only if $\pi(A, B, C):=I_{s}+z C\left(I_{n}-z A\right)^{-1} B=k(z)$. Please note that as for VARMA realizations, for a transfer function $k(z)$ there exist multiple state space realizations $(A, B, C)$, with possibly different state dimensions $n$. A state space system $(A, B, C)$ is minimal if there exists no state space system of lower state dimension realizing the same transfer function $k(z)$. The order of the transfer function $k(z)$ is the state dimension of a minimal system $(A, B, C)$ realizing $k(z)$.

All minimal state space realizations of a transfer function $k(z)$ only differ in the basis of the state (cf. Hannan and Deistler 1988, Theorem 2.3.4), i.e., $\pi(A, B, C)=\pi(\tilde{A}, \tilde{B}, \tilde{C})$ for two minimal state space systems $(A, B, C)$ and $(\tilde{A}, \tilde{B}, \tilde{C})$ is equivalent to the existence of a regular matrix $T \in \mathbb{C}^{n}$ such that $A=T \tilde{A} T^{-1}, B=T \tilde{B}, C=\tilde{C} T^{-1}$. Thus, the matrices $A$ and $\tilde{A}$ are similar for all minimal realizations of a transfer function $k(z)$.

By imposing restrictions on the matrices of a minimal state space system ( $A, B, C$ ) realizing $k(z)$, Bauer and Wagner (2012, Theorem 2) provide a canonical form, i.e., a mapping of the set $M_{n}$ of transfer functions with real valued power series coefficients defined below onto unique state space realizations $(\mathcal{A}, \mathcal{B}, \mathcal{C})$. The set $M_{n}$ is defined as

$$
M_{n}:=\left\{k(z)=\left.\pi(A, B, C)\right|_{A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times s}, C \in \mathbb{R}^{s \times n},(A, B, C) \text { minimal }} ^{\lambda_{\mid \max }(A) \leq 1}\right\}
$$

To describe the necessary restrictions of the canonical form the following definition is useful:

[^23]Definition 2. A matrix $B=\left[b_{i, j}\right]_{i=1, \ldots, c, j=1, \ldots, s} \in \mathbb{C}^{c \times s}$ is positive upper triangular (p.u.t.) if there exist integers $1 \leq j_{1} \leq j_{2} \leq \cdots \leq j_{c} \leq s$, such that for $j_{i} \leq s$ we have $b_{i, j}=0, j<j_{i}, j_{i}<j_{i+1}, b_{i, j_{i}} \in \mathbb{R}^{+}$; i.e., $B$ is of the form

$$
B=\left[\begin{array}{ccccccccc}
0 & \cdots & 0 & b_{1, j_{1}} & * & & \cdots & & * \\
0 & & \cdots & & 0 & b_{2, j_{2}} & * & & \\
0 & & & & & & 0 & b_{c, j_{c}} & *
\end{array}\right]
$$

where the symbol $*$ indicates unrestricted complex-valued entries.
A unique state space realization of $k(z) \in M_{n}$ is given as follows (cf. Bauer and Wagner 2012, Theorem 2):

Theorem 1. For every transfer function $k(z) \in M_{n}$ there exists a unique minimal (complex) state space realization $(\mathcal{A}, \mathcal{B}, \mathcal{C})$ such that

$$
\begin{aligned}
y_{t} & =\mathcal{C} x_{t, \mathbb{C}}+\Phi d_{t}+\varepsilon_{t} \\
x_{t+1, \mathbb{C}} & =\mathcal{A} x_{t, \mathbb{C}}+\mathcal{B} \varepsilon_{t}
\end{aligned}
$$

with:
(i) $\mathcal{A}:=\operatorname{diag}\left(\mathcal{A}_{u}, \mathcal{A}_{\bullet}\right):=\operatorname{diag}\left(\mathcal{A}_{1, \mathbb{C}}, \ldots, \mathcal{A}_{l, \mathbb{C}}, \mathcal{A}_{\bullet}\right), \mathcal{A}_{u} \in \mathbb{C}^{n_{u} \times n_{u}}, \mathcal{A} \bullet \in \mathbb{R}^{n_{\bullet} \times n_{\bullet}}$, where it holds for $k=1, \ldots, l$ that

- $\quad$ for $0<\omega_{k}<\pi$ :

$$
\mathcal{A}_{k, \mathbb{C}}:=\left[\begin{array}{cc}
J_{k} & 0 \\
0 & \bar{J}_{k}
\end{array}\right] \in \mathbb{C}^{2 d^{k} \times 2 d^{k}}
$$

- $\quad$ for $\omega_{k} \in\{0, \pi\}:$

$$
\mathcal{A}_{k, \mathbb{C}}:=\quad J_{k} \in \mathbb{R}^{d^{k} \times d^{k}}
$$

with

$$
J_{k}:=\left[\begin{array}{ccccc}
\overline{z_{k}} I_{d_{1}^{k}} & {\left[I_{d_{1}^{k}}, 0_{d_{1}^{k} \times\left(d_{2}^{k}-d_{1}^{k}\right)}\right]} & 0 & \cdots & 0  \tag{10}\\
0_{d_{2}^{k} \times d_{1}^{k}} & \overline{z_{k}} I_{d_{2}^{k}} & {\left[I_{d_{2}^{k}}, 0_{d_{2}^{k} \times\left(d_{3}^{k}-d_{2}^{k}\right)}\right]} & 0 & \vdots \\
0 & 0 & \overline{z_{k}} I_{d_{3}^{k}} & \ddots & 0 \\
\vdots & \vdots & \ddots & \ddots & {\left[I_{d_{h_{k}-1}^{k}}, 0_{d_{k_{k_{k}}^{k}} \times\left(d_{n_{k_{k}}}-d_{h_{k}-1}^{k}\right)}\right]} \\
0 & 0 & \cdots & 0 & \bar{z}_{k} I_{d_{h_{k}}}
\end{array}\right],
$$

where $0<d_{1}^{k} \leq d_{2}^{k} \leq \cdots \leq d_{h_{k}}^{k}$.
(ii) $\mathcal{B}:=\left[\mathcal{B}_{u}^{\prime}, \mathcal{B}_{\bullet}^{\prime}\right]^{\prime}:=\left[\mathcal{B}_{1, \mathbb{C}}^{\prime}, \ldots, \mathcal{B}_{l, \mathbb{C}}^{\prime}, \mathcal{B}_{\bullet}^{\prime}\right]^{\prime}$ and $\mathcal{C}:=\left[\mathcal{C}_{u}, \mathcal{C}_{\bullet}\right]:=\left[\mathcal{C}_{1, \mathbb{C}}, \ldots, \mathcal{C}_{l, \mathbb{C}}, \mathcal{C}_{\bullet}\right]$ are partitioned accordingly. It holds for $k=1, \ldots$, $l$ that

- $\quad$ for $0<\omega_{k}<\pi$ :

$$
\mathcal{B}_{k, \mathbb{C}}:=\left[\begin{array}{c}
\mathcal{B}_{k} \\
\overline{\mathcal{B}}_{k}
\end{array}\right] \in \mathbb{C}^{2 d^{k} \times s} \text { and } \mathcal{C}_{k, \mathbb{C}}:=\left[\mathcal{C}_{k}, \overline{\mathcal{C}}_{k}\right] \in \mathbb{C}^{s \times 2 d^{k}}
$$

- for $\omega_{k} \in\{0, \pi\}$ :

$$
\mathcal{B}_{k, \mathbb{C}}:=\mathcal{B}_{k} \in \mathbb{R}^{d^{k} \times s} \text { and } \mathcal{C}_{k, \mathbb{C}}:=\mathcal{C}_{k} \in \mathbb{R}^{s \times d^{k}}
$$

(iii) Partitioning $\mathcal{B}_{k, h_{k}}$ in $\mathcal{B}_{k}=\left[\mathcal{B}_{k, 1}^{\prime}, \ldots, \mathcal{B}_{k, h_{k}}^{\prime}\right]^{\prime}$ as $\mathcal{B}_{k, h_{k}}=\left[\mathcal{B}_{k, h_{k}, 1}^{\prime}, \ldots, \mathcal{B}_{k, h_{k}, h_{k}}^{\prime}\right]^{\prime}$, with $\mathcal{B}_{k, h_{k}, j} \in$ $\mathbb{C}^{\left(d_{j}^{k}-d_{j-1}^{k}\right) \times s}$ it holds that $\mathcal{B}_{k, h_{k}, j}$ is $p$. u.t. for $d_{j}^{k}>d_{j-1}^{k}$ for $j=1, \ldots, h_{k}$ and $k=1, \ldots, l$.
(iv) For $k=1, \ldots, l$ define $\mathcal{C}_{k}=\left[\mathcal{C}_{k, 1}, \mathcal{C}_{k, 2}, \ldots, \mathcal{C}_{k, h_{k}}\right], \mathcal{C}_{k, j}=\left[\mathcal{C}_{k, j}^{G}, \mathcal{C}_{k, j}^{E}\right]$, with $\mathcal{C}_{k, j}^{E} \in \mathbb{C}^{s \times\left(d_{j}^{k}-d_{j-1}^{k}\right)}$ and $\mathcal{C}_{k, j}^{G} \in \mathbb{C}^{s \times d_{j-1}^{k}}$ for $j=1, \ldots, h_{k}$, with $d_{0}^{k}:=0$. Furthermore, define $\mathcal{C}_{k}^{E}:=\left[\mathcal{C}_{k, 1}^{E}, \ldots, \mathcal{C}_{k, h_{k}}^{E}\right] \in \mathbb{C}^{s \times d_{h_{k}}^{k}}$. It holds that $\left(\mathcal{C}_{k}^{E}\right)^{\prime} \mathcal{C}_{k}^{E}=I_{d_{h_{k}}}$ and $\left(\mathcal{C}_{k, j}^{G}\right)^{\prime} \mathcal{C}_{k, i}^{E}=0$ for $1 \leq i \leq j$ for $j=2, \ldots, h_{k}$ and $k=1, \ldots, l$.
(v) $\lambda_{|\max |}\left(\mathcal{A}_{\bullet}\right)<1$ and the stable subsystem $\left(\mathcal{A}_{\bullet}, \mathcal{B}_{\bullet}, \mathcal{C}_{\bullet}\right)$ of state dimension $n_{\bullet}=n-n_{u}$ is in echelon canonical form (cf. Hannan and Deistler 1988, Theorem 2.5.2).

Remark 4. As indicated in Remark 3 and discussed in detail in Bauer and Wagner (2012) considering complex valued quantities is merely for algebraic convenience. For econometric analysis, interest is, of course, on real valued quantities. These can be straightforwardly obtained from the representation given in Theorem 1 as follows. First define a transformation matrix (and its inverse):

$$
T_{\mathbb{R}, d}:=\left[I_{d} \otimes\left[\begin{array}{l}
1 \\
i
\end{array}\right], I_{d} \otimes\left[\begin{array}{c}
1 \\
-i
\end{array}\right]\right] \in \mathbb{C}^{2 d \times 2 d}, \quad T_{\mathbb{R}, d}^{-1}:=\frac{1}{2}\left[\begin{array}{c}
I_{d} \otimes[1,-i] \\
I_{d} \otimes[1,
\end{array}\right] .
$$

Starting from the complex valued canonical representation $(\mathcal{A}, \mathcal{B}, \mathcal{C})$, a real valued canonical representation

$$
\begin{aligned}
y_{t} & =\mathcal{C}_{\mathbb{R}} x_{t, \mathbb{R}}+\Phi d_{t}+\varepsilon_{t}, \\
x_{t+1, \mathbb{R}} & =\mathcal{A}_{\mathbb{R}} x_{t, \mathbb{R}}+\mathcal{B}_{\mathbb{R}} \varepsilon_{t},
\end{aligned}
$$

with real valued matrices $\left(\mathcal{A}_{\mathbb{R}}, \mathcal{B}_{\mathbb{R}}, \mathcal{C}_{\mathbb{R}}\right)$ follows from using the just defined transformation matrix. In particular it holds that:

$$
\begin{aligned}
& \mathcal{A}_{\mathbb{R}}:=\operatorname{diag}\left(\mathcal{A}_{u, \mathbb{R}}, \mathcal{A}_{\bullet}\right):=\operatorname{diag}\left(\mathcal{A}_{1, \mathbb{R}}, \ldots, \mathcal{A}_{l, \mathbb{R}}, \mathcal{A}_{\bullet}\right), \\
& \mathcal{B}_{\mathbb{R}}:=\left[\mathcal{B}_{u, \mathbb{R}^{\prime}}^{\prime}, \mathcal{B}_{\bullet}^{\prime}\right]^{\prime} \quad:=\left[\mathcal{B}_{1, \mathbb{R}}^{\prime}, \ldots, \mathcal{B}_{l, \mathbb{R}}^{\prime}, \mathcal{B}_{\bullet}^{\prime}\right]^{\prime}, \\
& \mathcal{C}_{\mathbb{R}}:=\quad\left[\mathcal{C}_{u, \mathbb{R}}, \mathcal{C}_{\bullet}\right] \quad:=\left[\mathcal{C}_{1, \mathbb{R}}, \ldots, \mathcal{C}_{l, \mathbb{R}}, \mathcal{C}_{\bullet}\right],
\end{aligned}
$$

with

$$
\left(\mathcal{A}_{k, \mathbb{R}}, \mathcal{B}_{k, \mathbb{R}}, \mathcal{C}_{k, \mathbb{R}}\right):= \begin{cases}\left(T_{\mathbb{R}, d^{k}} \mathcal{A}_{k} T_{\mathbb{R}, d^{k}}^{-1}, T_{\mathbb{R}, d^{k}} \mathcal{B}_{k}, \mathcal{C}_{k} T_{\mathbb{R}, d^{k}}^{-1}\right) & \text { if } 0<\omega_{k}<\pi \\ \left(\mathcal{A}_{k}, \mathcal{B}_{k}, \mathcal{C}_{k}\right) & \text { if } \omega_{k} \in\{0, \pi\}\end{cases}
$$

Before we turn to the real valued state process corresponding to the real valued canonical representation, we first consider the complex valued state process $\left\{x_{t, \mathbb{C}}\right\}_{t \in \mathbb{Z}}$ in more detail. This process is partitioned according to the partitioning of the matrices $\mathcal{C}_{k, \mathbb{C}}$ into $x_{t, \mathbb{C}}:=\left[x_{t, u}^{\prime}, x_{t, \bullet}^{\prime}\right]^{\prime}:=\left[x_{t, 1, \mathbb{C}}^{\prime}, \ldots, x_{t, l, \mathbb{C}}^{\prime}, x_{t, \bullet}^{\prime}\right]^{\prime}$, where

$$
x_{t, k, \mathbb{C}}:= \begin{cases}{\left[x_{t, k}^{\prime} \bar{x}_{t, k}^{\prime}\right]^{\prime}} & \text { if } 0<\omega_{k}<\pi \\ x_{t, k} & \text { if } \omega_{k} \in\{0, \pi\}\end{cases}
$$

with

$$
x_{t+1, k}=J_{k} x_{t, k}+\mathcal{B}_{k} \varepsilon_{t}, \quad \text { for } k=1, \ldots, l .
$$

For $k=1, \ldots, l$ the sub-vectors $x_{t, k}$ are further decomposed into $x_{t, k}:=\left[\left(x_{t, k}^{1}\right)^{\prime}, \ldots,\left(x_{t, k}^{h_{k}}\right)^{\prime}\right]^{\prime}$, with $x_{t, k}^{j} \in \mathbb{C}^{d_{j}^{k}}$ for $j=1, \ldots, h_{k}$ according to the partitioning $\mathcal{C}_{k}=\left[\mathcal{C}_{k, 1}, \ldots, \mathcal{C}_{k, h_{k}}\right]$.

The partitioning of the complex valued process $\left\{x_{t, \mathbb{C}}\right\}_{t \in \mathbb{Z}}$ leads to an analogous partitioning of the real valued state process $\left\{x_{t, \mathbb{R}}\right\}_{t \in \mathbb{Z}}, x_{t, \mathbb{R}}:=\left[x_{t, u, \mathbb{R}^{\prime}}^{\prime}, x_{t, \bullet}^{\prime}\right]^{\prime}:=\left[x_{t, 1, \mathbb{R}^{\prime}}^{\prime} \ldots, x_{t, l, \mathbb{R}}^{\prime}, x_{t, \bullet}^{\prime}\right]^{\prime}$, obtained from

$$
x_{t, k, \mathbb{R}}:= \begin{cases}T_{\mathbb{R}, d^{k}} x_{t, k, \mathbb{C}} & \text { if } 0<\omega_{k}<\pi \\ x_{t, k} & \text { if } \omega_{k} \in\{0, \pi\}\end{cases}
$$

with the corresponding block of the state equation given by

$$
x_{t+1, k, \mathbb{R}}=\mathcal{A}_{k, \mathbb{R}} x_{t, k, \mathbb{R}}+\mathcal{B}_{k, \mathbb{R}^{\varepsilon_{t}}} .
$$

For $k=1, \ldots, l$ the sub-vectors $x_{t, k, \mathbb{R}}$ are further decomposed into $x_{t, k, \mathbb{R}}:=\left[\left(x_{t, k, \mathbb{R}}^{1}\right)^{\prime}, \ldots,\left(x_{t, k, \mathbb{R}}^{h_{k}}\right)^{\prime}\right]^{\prime}$, with $x_{t, k, \mathbb{R}}^{j} \in \mathbb{R}^{2 d_{j}^{k}}$ if $0<\omega_{k}<\pi$ and $x_{t, k, \mathbb{R}}^{j} \in \mathbb{R}^{d_{j}^{k}}$ if $\omega_{k} \in\{0, \pi\}$ for $j=1, \ldots, h_{k}$ and $\mathcal{C}_{k, \mathbb{R}}:=$ $\left[\mathcal{C}_{k, 1, \mathbb{R}}, \ldots, \mathcal{C}_{k, h_{k}, \mathbb{R}}\right]$ decomposed accordingly.

Bauer and Wagner (2012, Theorem 3, p. 1328) show that the processes $\left\{x_{t, k, \mathbb{R}}^{j}\right\}_{t \in \mathbb{Z}}$ have unit root structure $\left(\left(\omega_{k}, h_{k}-j+1\right)\right)$ for $j=1, \ldots, h_{k}$ and $k=1, \ldots, l$. Furthermore, for $j=1, \ldots, h_{k}$ and $k=1, \ldots, l$ the processes $\left\{x_{t, k, \mathbb{R}}^{j}\right\}_{t \in \mathbb{Z}}$ are not cointegrated, as defined in Definition 3 below. For $\omega_{k}=0$, the process $\left\{x_{t, k, \mathbb{R}}^{j}\right\}_{t \in \mathbb{Z}}$ is the $d_{k}^{j}$-dimensional process of stochastic trends of order $h_{1}-j+1$, while the $2 d_{j}^{k}$ components of $\left\{x_{t, k, \mathbb{R}}^{j}\right\}_{t \in \mathbb{Z}}$, for $0<\omega_{k}<\pi$, and the $d_{j}^{k}$ components of $\left\{x_{t, l, \mathbb{R}}^{j}\right\}_{t \in \mathbb{Z}}$, for $\omega_{k}=\pi$, are referred to as stochastic cycles of order $h_{k}-j+1$ at their corresponding frequencies $\omega_{k}$.

Remark 5. Parameterizing the stable part of the transfer function using the echelon canonical form is merely one possible choice. Any other canonical form of the stable subsystem and suitable parameterization based on it can be used instead for the stable subsystem.

Remark 6. Starting from a state space system (9) with matrices $(\mathcal{A}, \mathcal{B}, \mathcal{C})$ in canonical form, a solution for $y_{t}, t>0$ (with the solution for $t<0$ obtained completely analogously) -for some $x_{1}=\left[x_{1, u}^{\prime}, x_{1, \bullet}^{\prime}\right]^{\prime}$ —is given by

$$
y_{t}=\sum_{j=1}^{t-1} \mathcal{C}_{u} \mathcal{A}_{u}^{j-1} \mathcal{B}_{u} \varepsilon_{t-j}+\mathcal{C}_{u} \mathcal{A}_{u}^{t-1} x_{1, u}+\sum_{j=1}^{t-1} \mathcal{C}_{\bullet} \mathcal{A}_{\bullet}^{j-1} \mathcal{B}_{\bullet} \varepsilon_{t-j}+\mathcal{C}_{\bullet} \mathcal{A}_{\bullet}^{t-1} x_{1, \bullet}+\Phi d_{t}+\varepsilon_{t}
$$

Clearly, the term $\mathcal{C}_{u} \mathcal{A}_{u}^{t-1} x_{1, u}$ is stochastically singular and is effectively like a deterministic component, which may lead to an identification problem with $\Phi d_{t}$. If, the deterministic component $\Phi d_{t}$ is rich enough to "absorb" $\mathcal{C}_{u} \mathcal{A}_{u}^{t-1} x_{1, u}$, then one solution of the identification problem is to set $x_{1, u}=0$. Rich enough here means, e.g., in the $I(1)$ case with $\mathcal{A}_{u}=I$ that $d_{t}$ contains an intercept. Analogously, in the MFI(1) case $d_{t}$ has to contain seasonal dummy variables corresponding to all unit root frequencies. The term $\mathcal{C} \cdot \mathcal{A}_{\bullet}^{t-1} x_{1, \bullet}$ decays exponentially and, therefore, does not impact the asymptotic properties of any statistical procedure. It is, therefore, inconsequential for statistical analysis but convenient (with respect to our definition of unit root processes) to set $x_{1, \bullet}=\sum_{j=1}^{\infty} \mathcal{A}_{\bullet}^{j-1} \mathcal{B} \bullet \varepsilon_{1-j}$. This corresponds to the steady state or stationary solution of the stable block of the state equation, and renders $\left\{x_{t, \bullet}\right\}_{t \in \mathbb{N}}$ or, when the solution on $\mathbb{Z}$ is considered, $\left\{x_{t, \bullet}\right\}_{t \in \mathbb{Z}}$ stationary. Please note that these issues with respect to starting values, potential identification problems and their impact or non-impact on statistical procedures also occur in the VAR setting.

Bauer and Wagner (2012, Theorem 2) show that minimality of the canonical state space realization $(\mathcal{A}, \mathcal{B}, \mathcal{C})$ implies full row rank of the p.u.t. blocks $\mathcal{B}_{k, h_{k}, j}$ of $\mathcal{B}_{k, h_{k}}$. In addition to proposing the canonical form, Bauer and Wagner (2012) also provide details how to transform any minimal state space realization into canonical form: Given a minimal state space system $(A, B, C)$ realizing the transfer function $k(z) \in M_{n}$, the first step is to find a similarity transformation $T$ such that $\tilde{A}=T A T^{-1}$
is of the form given in (10) by using an eigenvalue decomposition, compare Chatelin (1993). In the second step the corresponding subsystem $\left(\tilde{A}_{\bullet}, \tilde{B}_{\bullet}, \tilde{C}_{\bullet}\right)$ is transformed to echelon canonical form as described in Hannan and Deistler (1988, chp. 2). These two transformations do not lead to a unique realization, because the restrictions on $\mathcal{A}$ do not uniquely determine the unstable subsystem $\left(\mathcal{A}_{u}, \mathcal{B}_{u}, \mathcal{C}_{u}\right)$.

For example, in the case $\Omega=\left(\left(\omega_{1}, h_{1}\right)\right)=((0,1)), n_{\bullet}=0, d_{1}^{1}<s$, such that $\left(I_{d_{1}^{1}}, \mathcal{B}_{1}, \mathcal{C}_{1}\right)$ is a corresponding state space system, the same transfer function $k(z)=I_{s}+z \mathcal{C}_{1}(1-z)^{-1} \mathcal{B}_{1}=I_{s}+$ $\mathcal{C}_{1} \mathcal{B}_{1} z(1-z)^{-1}$ is realized also by all systems $\left(I_{d_{1}^{1}}, T \mathcal{B}_{1}, \mathcal{C}_{1} T^{-1}\right)$, with some regular matrix $T \in \mathbb{C}_{1}^{d_{1}^{1} \times d_{1}^{1}}$. To find a unique realization the product $\mathcal{C}_{1} \mathcal{B}_{1}$ needs to be uniquely decomposed into factors $\mathcal{C}_{1}$ and $\mathcal{B}_{1}$. This is achieved by performing a QR decomposition of $\mathcal{C}_{1} \mathcal{B}_{1}$ (without pivoting) that leads to $\mathcal{C}_{1}^{\prime} \mathcal{C}_{1}=I$. The additional restriction of $\mathcal{B}_{1}$ being a p.u.t. matrix of full row rank then leads to a unique factorization of $\mathcal{C}_{1} \mathcal{B}_{1}$ into $\mathcal{C}_{1}$ and $\mathcal{B}_{1}$. In the general case with an arbitrary unit root structure $\Omega$, similar arguments lead to p.u.t. restrictions on sub-blocks $\mathcal{B}_{k, h_{k}, j}$ in $\mathcal{B}_{u}$ and orthogonality restrictions on sub-blocks of $\mathcal{C}_{u}$.

The canonical form introduced in Theorem 1 was designed to be useful for cointegration analysis. To see this, first requires a definition of static and polynomial cointegration (cf. Bauer and Wagner 2012, Definitions 3 and 4).

## Definition 3.

(i) Let $\tilde{\Omega}=\left(\left(\tilde{\omega}_{1}, \tilde{h}_{1}\right), \ldots,\left(\tilde{\omega}_{\tilde{l}}, \tilde{h}_{\tilde{l}}\right)\right)$ and $\Omega=\left(\left(\omega_{1}, h_{1}\right), \ldots,\left(\omega_{l}, h_{l}\right)\right)$ be two unit root structures. Then $\tilde{\Omega} \preceq$ $\Omega$ if

- $\quad F(\tilde{\Omega}) \subseteq F(\Omega)$.
- For all $\omega \in F(\tilde{\Omega})$ for $\tilde{k}$ and $k$ such that $\tilde{\omega}_{\tilde{k}}=\omega_{k}=\omega$ it holds that $\tilde{h}_{\tilde{k}} \leq h_{k}$.

Furthermore, $\tilde{\Omega} \prec \Omega$ if $\tilde{\Omega} \preceq \Omega$ and $\tilde{\Omega} \neq \Omega$. For two unit root structures $\tilde{\Omega} \preceq \Omega$ define the decrease $\delta_{k}(\Omega, \tilde{\Omega})$ of the integration order at frequency $\omega_{k}$, for $k=1, \ldots, l$, as

$$
\delta_{k}(\Omega, \tilde{\Omega}):= \begin{cases}h_{k}-\tilde{h}_{\tilde{k}} & \exists \tilde{k}: \tilde{\omega}_{\tilde{k}}=\omega_{k} \in F(\tilde{\Omega}) \\ h_{k} & \omega_{k} \notin F(\tilde{\Omega})\end{cases}
$$

(ii) An s-dimensional unit root process $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ with unit root structure $\Omega$ is cointegrated of order $(\Omega, \tilde{\Omega})$, where $\tilde{\Omega} \prec \Omega$, if there exists a vector $\beta \in \mathbb{R}^{s}, \beta \neq 0$, such that $\left\{\beta^{\prime} y_{t}\right\}_{t \in \mathbb{Z}}$ has unit root structure $\tilde{\Omega}$. In this case the vector $\beta$ is a cointegrating vector (CIV) of order $(\Omega, \tilde{\Omega})$.
(iii) All CIVs of order $(\Omega, \tilde{\Omega})$ span the (static) cointegrating space of order $(\Omega, \tilde{\Omega}) .{ }^{16}$
(iv) An s-dimensional unit root process $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ with unit root structure $\Omega$ is polynomially cointegrated of order $(\Omega, \tilde{\Omega})$, where $\tilde{\Omega} \prec \Omega$, if there exists a vector polynomial $\beta(z)=\sum_{m=0}^{q} \beta_{m} z^{m}, \beta_{m} \in \mathbb{R}^{s}, m=$ $0, \ldots, q, \beta_{q} \neq 0$, for some integer $1 \leq q<\infty$ such that

- $\quad \beta(L)^{\prime}\left(\left\{y_{t}\right\}_{t \in \mathbb{Z}}\right)$ has unit root structure $\tilde{\Omega}$,
- $\max _{k=1, \ldots, l}\left\|\beta\left(e^{i \omega_{k}}\right)\right\| \delta_{k}(\Omega, \tilde{\Omega}) \neq 0$.

In this case the vector polynomial $\beta(z)$ is a polynomial cointegrating vector (PCIV) of order $(\Omega, \tilde{\Omega})$.
(v) All PCIVs of order $(\Omega, \tilde{\Omega})$ span the polynomial cointegrating space of order $(\Omega, \tilde{\Omega})$.

[^24]
## Remark 7.

(i) It is merely a matter of taste whether cointegrating spaces are defined in terms of their order $(\Omega, \tilde{\Omega})$ or their decrease $\delta(\Omega, \tilde{\Omega}):=\left(\delta_{1}(\Omega, \tilde{\Omega}), \ldots, \delta_{l}(\Omega, \tilde{\Omega})\right)$, with $\delta_{k}(\Omega, \tilde{\Omega})$ as defined above. Specifying $\Omega$ and $\delta(\Omega, \tilde{\Omega})$ contains the same information as providing the order of (polynomial) cointegration.
(ii) Notwithstanding the fact that CIVs and PCIVs in general may lead to changes of the integration orders at different unit root frequencies it may be of interest to "zoom in" on only one unit root frequency $\omega_{k}$, thereby leaving the potential reductions of the integration orders at other unit root frequencies unspecified. This allows to—entirely similarly as in Definition 3—define cointegrating and polynomial cointegrating spaces of different orders at a single unit root frequency $\omega_{k}$. Analogously one can also define cointegrating and polynomial cointegrating spaces of different orders for subsets of the frequencies in $F(\Omega)$.
(iii) In principle the polynomial cointegrating spaces defined so far are infinite-dimensional as the polynomial degree is not bounded. However, since every polynomial vector $\beta(z)$ can be written as $\beta_{0}(z)+\beta_{\Omega}(z) \Delta_{\Omega}(z)$, where by definition $\left\{\Delta_{\Omega} y_{t}\right\}_{t \in \mathbb{Z}}$ has empty unit root structure, it suffices to consider PCIVs of polynomial degree smaller than the polynomial degree of $\Delta_{\Omega}(z)$. This shows that it is sufficient to consider finite dimensional polynomial cointegrating spaces. When considering, as in item (ii), (polynomial) cointegration only for one unit root it similarly suffices to consider polynomials of maximal degree equal to $h_{k}-1$ for real unit roots and $2 h_{k}-1$ for complex unit roots. Thus, in the $I(2)$ case it suffices to consider polynomials of degree one.
(iv) The argument about maximal relevant polynomial degrees given in item (iii) can be made more precise and combined with the decrease in $\Omega$ achieved. Every polynomial vector $\beta(z)$ can be written as $\beta_{0}(z)+$ $\beta_{\omega_{k}, \delta_{k}}(z) \Delta_{\omega_{k}}^{\delta_{k}}(z)$ for $\delta_{k}=1, \ldots, h_{k}$. By definition it holds that $\left\{\Delta_{\omega_{k}}^{\delta_{k}} y_{t}\right\}_{t \in \mathbb{Z}}$ has integration order $h_{k}-\delta_{k}$ at frequency $\omega_{k}$. Thus, it suffices to consider PCIVs of polynomial degree smaller than $\delta_{k}$ for $\omega_{k} \in\{0, \pi\}$ or $2 \delta_{k}$ for $0<\omega_{k}<\pi$ when considering the polynomial cointegrating space at $\omega_{k}$ with decrease $\delta_{k}$. In the MFI(1) case therefore, when considering only one unit root frequency, again only polynomials of degree one need to be considered. This space is often referred to in the literature as dynamic cointegration space.

To illustrate the advantages of the canonical form for cointegration analysis consider

$$
y_{t}=\sum_{k=1}^{l} \sum_{j=1}^{h_{k}} \mathcal{C}_{k, j, \mathbb{R}} x_{t, k, \mathbb{R}}^{j}+\mathcal{C}_{\bullet} x_{t, \bullet}+\Phi d_{t}+\varepsilon_{t}
$$

By Remark 4, the process $\left\{x_{t, k, \mathbb{R}}^{j}\right\}_{t \in \mathbb{Z}}$ is not cointegrated. This implies that $\beta \in \mathbb{R}^{s}, \beta \neq 0$, reduces the integration order at unit root $z_{k}$ to $h_{k}-j$ if and only if $\beta^{\prime}\left[\mathcal{C}_{k, 1, \mathbb{R}}, \ldots, \mathcal{C}_{k, j, \mathbb{R}}\right]=0$ and $\beta^{\prime} \mathcal{C}_{k, j+1, \mathbb{R}} \neq 0$ or equivalently $\beta^{\prime}\left[\mathcal{C}_{k, 1}, \ldots, \mathcal{C}_{k, j}\right]=0$ and $\beta^{\prime} \mathcal{C}_{k, j+1} \neq 0$ (using the transformation to the complex matrices of the canonical form, as discussed in Remark 4, and that $\beta^{\prime}\left[\mathcal{C}_{k}, \overline{\mathcal{C}_{k}}\right]=0$ if and only if $\left.\beta^{\prime} \mathcal{C}_{k}=0\right)$. Thus, the CIVs are characterized by orthogonality to sub-blocks of $\mathcal{C}_{u}$.

The real valued representation given in Remark 4 used in its partitioned form just above immediately leads to necessary orthogonality constraint for polynomial cointegration of degree one:

$$
\begin{aligned}
\beta(L)^{\prime}\left(y_{t}\right) & =\beta(L)^{\prime}\left(\mathcal{C}_{u, \mathbb{R}} x_{t, u, \mathbb{R}}+\mathcal{C} \bullet x_{t, \bullet}+\Phi d_{t}+\varepsilon_{t}\right) \\
& =\beta_{0}^{\prime} \mathcal{C}_{u, \mathbb{R}} x_{t, u, \mathbb{R}}+\beta_{1}^{\prime} \mathcal{C}_{u, \mathbb{R}} x_{t-1, u, \mathbb{R}}+\beta(L)^{\prime}\left(\mathcal{C}_{\bullet} x_{t, \bullet}+\Phi d_{t}+\varepsilon_{t}\right) \\
& =\beta_{0}^{\prime} \mathcal{C}_{u, \mathbb{R}}\left(\mathcal{A}_{u, \mathbb{R}} x_{t-1, u, \mathbb{R}}+\mathcal{B}_{u, \mathbb{R}} \varepsilon_{t-1}\right)+\beta_{1}^{\prime} \mathcal{C}_{u, \mathbb{R}} x_{t-1, u, \mathbb{R}}+\beta(L)^{\prime}\left(\mathcal{C}_{\bullet} x_{t, \bullet}+\Phi d_{t}+\varepsilon_{t}\right) \\
& =\left(\beta_{0}^{\prime} \mathcal{C}_{u, \mathbb{R}} \mathcal{A}_{u, \mathbb{R}}+\beta_{1}^{\prime} \mathcal{C}_{u, \mathbb{R}}\right) x_{t-1, u, \mathbb{R}}+\beta_{0}^{\prime} \mathcal{C}_{u, \mathbb{R}} \mathcal{B}_{u, \mathbb{R}} \varepsilon_{t-1}+\beta(L)^{\prime}\left(\mathcal{C}_{\bullet} x_{t, \bullet}+\Phi d_{t}+\varepsilon_{t}\right) \\
& =\left(\beta_{0}^{\prime} \mathcal{C}_{u} \mathcal{A}_{u}+\beta_{1}^{\prime} \mathcal{C}_{u}\right) x_{t-1, u}+\beta_{0}^{\prime} \mathcal{C}_{u} \mathcal{B}_{u} \varepsilon_{t-1}+\beta(L)^{\prime}\left(\mathcal{C}_{\bullet} x_{t, \bullet}+\Phi d_{t}+\varepsilon_{t}\right)
\end{aligned}
$$

follows. Since all terms except the first are stationary or deterministic, a necessary condition for a reduction of the unit root structure is the orthogonality of $\left[\begin{array}{lll}\beta_{0}^{\prime} & \beta_{1}^{\prime}\end{array}\right]^{\prime}$ to sub-blocks of $\left[\begin{array}{c}\mathcal{C}_{u}, \mathbb{R} \\ \mathcal{A}_{u, \mathbb{R}}\end{array}\right]$ or
sub-blocks of the complex matrix $\left[\begin{array}{c}\mathcal{C}_{u} \mathcal{A}_{u} \\ \mathcal{C}_{u}\end{array}\right]$. Please note, however, that this orthogonality condition is not sufficient for $\left[\beta_{0}^{\prime}, \beta_{1}^{\prime}\right]^{\prime}$ to be a PCIV, because it does not imply $\max _{k=1, \ldots, l}\left\|\beta\left(e^{i \omega_{k}}\right)\right\| \delta_{k}(\Omega, \tilde{\Omega}) \neq 0$. For a detailed discussion of polynomial cointegration, when considering also higher polynomial degrees, see Bauer and Wagner (2012, sct. 5).

The following examples illustrate cointegration analysis in the state space framework for the empirically most relevant, i.e., the $\mathrm{I}(1), \mathrm{MFI}(1)$ and $\mathrm{I}(2)$ cases.

Example 1 (Cointegration in the $\mathrm{I}(1)$ case). In the $\mathrm{I}(1)$ case, neglecting the stable subsystem and the deterministic components for simplicity, it holds that

$$
\begin{aligned}
y_{t} & =\mathcal{C}_{1} x_{t, 1}+\varepsilon_{t}, \quad y_{t}, \varepsilon_{t} \in \mathbb{R}^{s}, x_{t, 1} \in \mathbb{R}^{d_{1}^{1}}, \mathcal{C}_{1} \in \mathbb{R}^{s \times d_{1}^{1}} \\
x_{t+1,1} & =x_{t, 1}+\mathcal{B}_{1} \varepsilon_{t}, \quad \mathcal{B}_{1} \in \mathbb{R}^{d_{1}^{1} \times s} .
\end{aligned}
$$

The vector $\beta \in \mathbb{R}^{s}, \beta \neq 0$, is a CIV of order $((0,1),\{ \})$ if and only if $\beta^{\prime} \mathcal{C}_{1}=0$.
Example 2 (Cointegration in the MFI(1) case with complex unit root $z_{k}$ ). In the MFI(1) case with unit root structure $\Omega=\left(\left(\omega_{k}, 1\right)\right)$ and complex unit root $z_{k}$, neglecting the stable subsystem and the deterministic components for simplicity, it holds that

$$
\begin{aligned}
y_{t}= & \mathcal{C}_{k, \mathbb{R} x_{t, k, \mathbb{R}}+\varepsilon_{t}}= \\
= & {\left[\begin{array}{ll}
\mathcal{C}_{k} & \overline{\mathcal{C}}_{k}
\end{array}\right]\left[\begin{array}{l}
x_{t, k} \\
\bar{x}_{t, k}
\end{array}\right]+\varepsilon_{t}, } \\
& y_{t}, \varepsilon_{t} \in \mathbb{R}^{s}, x_{t, k, \mathbb{R}} \in \mathbb{R}^{2 d_{1}^{k}}, x_{t, k} \in \mathbb{C}_{1}^{d_{1}^{k}}, \mathcal{C}_{k, \mathbb{R}} \in \mathbb{R}^{s \times 2 d_{1}^{k}}, \mathcal{C}_{k} \in \mathbb{C}^{s \times d_{1}^{k}}, \\
{\left[\begin{array}{c}
x_{t+1, k} \\
\bar{x}_{t+1, k}
\end{array}\right]=} & {\left[\begin{array}{cc}
\bar{z}_{k} I_{d_{1}^{k}} & 0 \\
0 & z_{k} I_{d_{1}^{k}}
\end{array}\right]\left[\begin{array}{c}
x_{t, k} \\
\bar{x}_{t, k}
\end{array}\right]+\left[\begin{array}{c}
\mathcal{B}_{k} \\
\overline{\mathcal{B}}_{k}
\end{array}\right] \varepsilon_{t}, \quad \mathcal{B}_{k} \in \mathbb{C}^{d_{1}^{k} \times s} . }
\end{aligned}
$$

The vector $\beta \in \mathbb{R}^{s}, \beta \neq 0$, is a CIV of order $(\Omega,\{ \})$ if and only if

$$
\beta^{\prime} \mathcal{C}_{k}=0\left(\text { and thus } \beta^{\prime} \overline{\mathcal{C}}_{k}=0\right)
$$

The vector polynomial $\beta(z)=\beta_{0}+\beta_{1} z$, with $\beta_{0}, \beta_{1} \in \mathbb{R}^{s},\left[\beta_{0}^{\prime}, \beta_{1}^{\prime}\right]^{\prime} \neq 0$, is a PCIV of order $(\Omega,\{ \})$ if and only if

$$
\left[\beta_{0}^{\prime}, \beta_{1}^{\prime}\right]\left[\begin{array}{cc}
\bar{z}_{k} \mathcal{C}_{k} & z_{k} \overline{\mathcal{C}}_{k}  \tag{11}\\
\mathcal{C}_{k} & \overline{\mathcal{C}}_{k}
\end{array}\right]=0
$$

which is equivalent to

$$
\left(\bar{z}_{k} \beta_{0}^{\prime}+\beta_{1}^{\prime}\right) \mathcal{C}_{k}=0 .
$$

The fact that the matrix in (11) has a block structure with two blocks of conjugate complex columns implies some additional structure also on the space of PCIVs, here with polynomial degree one. More specifically it holds that if $\beta_{0}+\beta_{1} z$ is a PCIV of order $(\Omega,\{ \})$, also $-\beta_{1}+\left(\beta_{0}+2 \cos \left(\omega_{k}\right) \beta_{1}\right) z$ is a PCIV of order $(\Omega,\{ \})$. This follows from

$$
\begin{aligned}
\left(\bar{z}_{k}\left(-\beta_{1}\right)^{\prime}+\left(\beta_{0}+2 \cos \left(\omega_{k}\right) \beta_{1}\right)^{\prime}\right) \mathcal{C}_{k} & =\left(\beta_{0}^{\prime}+\left(2 \mathcal{R}\left(z_{k}\right)-\bar{z}_{k}\right) \beta_{1}^{\prime}\right) \mathcal{C}_{k} \\
& =\left(\beta_{0}^{\prime}+z_{k} \beta_{1}^{\prime}\right) \mathcal{C}_{k} \\
& =z_{k}\left(\bar{z}_{k} \beta_{0}^{\prime}+\beta_{1}^{\prime}\right) \mathcal{C}_{k}=0
\end{aligned}
$$

Thus, the space of PCIVs of degree (up to) one inherits some additional structure emanating from the occurrence of complex eigenvalues in complex conjugate pairs.

Example 3 (Cointegration in the $\mathrm{I}(2)$ case). In the $I(2)$ case, neglecting the stable subsystem and the deterministic components for simplicity, it holds that

$$
\begin{aligned}
y_{t}= & \mathcal{C}_{1,1}^{E} x_{t, 1}^{E}+\mathcal{C}_{1,2}^{G} x_{t, 2}^{G}+\mathcal{C}_{1,2}^{E} x_{t, 2}^{E}+\varepsilon_{t}, \\
& y_{t}, \varepsilon_{t} \in \mathbb{R}^{s}, x_{t, 1}^{E} x_{t, 2}^{G} \in \mathbb{R}^{d_{1}^{1}, x_{t, 2}^{E} \in \mathbb{R}^{d_{2}^{1}-d_{1}^{1}}, \mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{G} \in \mathbb{R}^{s \times d_{1}^{1}}, \mathcal{C}_{1,2}^{E} \in \mathbb{R}^{s \times\left(d_{2}^{1}-d_{1}^{1}\right)},} \\
x_{t+1,1}^{E}= & x_{t, 1}^{E}+x_{t, 2}^{G}+\mathcal{B}_{1,1} \varepsilon_{t}, \\
x_{t+1,2}^{G}= & x_{t, 2}^{G}+\mathcal{B}_{1,2,1} \varepsilon_{t}, \\
x_{t+1,2}^{E}= & x_{t, 2}^{E}+\mathcal{B}_{1,2,2} \varepsilon_{t}, \quad \mathcal{B}_{1,1} \in \mathbb{R}^{d_{1}^{1} \times s}, \mathcal{B}_{1,2,1} \in \mathbb{R}^{d_{1}^{1} \times s}, \mathcal{B}_{1,2,2} \in \mathbb{R}^{\left(d_{2}^{1}-d_{1}^{1}\right) \times s} .
\end{aligned}
$$

The vector $\beta \in \mathbb{R}^{s}, \beta \neq 0$ is a CIV of order $((0,2),(0,1))$ if and only if

$$
\beta^{\prime} \mathcal{C}_{1,1}^{E}=0 \quad \text { and } \quad \beta^{\prime}\left[\mathcal{C}_{1,2}^{G}, \mathcal{C}_{1,2}^{E}\right] \neq 0
$$

The vector $\beta \in \mathbb{R}^{s}, \beta \neq 0$, is a CIV of order $((0,2),\{ \})$ if and only if

$$
\beta^{\prime}\left[\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{G}, \mathcal{C}_{1,2}^{E}\right]=0
$$

The vector polynomial $\beta(z)=\beta_{0}+\beta_{1} z$, with $\beta_{0}, \beta_{1} \in \mathbb{R}^{s}$ is a PCIV of order $((0,2),\{ \})$ if and only if

$$
\left[\beta_{0}^{\prime}, \beta_{1}^{\prime}\right]\left[\begin{array}{ccc}
\mathcal{C}_{1,1}^{E} & \mathcal{C}_{1,1}^{E}+\mathcal{C}_{1,2}^{G} & \mathcal{C}_{1,2}^{E} \\
\mathcal{C}_{1,1}^{E} & \mathcal{C}_{1,2}^{G} & \mathcal{C}_{1,2}^{E}
\end{array}\right]=0 \quad \text { and } \quad \beta(1)=\beta_{0}+\beta_{1} \neq 0
$$

The above orthogonality constraint indicates that the two cases $\mathcal{C}_{1,2}^{G}=0$ and $\mathcal{C}_{1,2}^{G} \neq 0$ have to be considered separately for polynomial cointegration analysis. Consider first the case $\mathcal{C}_{1,2}^{G}=0$. In this case the orthogonality constraints imply $\beta_{0}^{\prime} \mathcal{C}_{1,1}^{E}=0, \beta_{1}^{\prime} \mathcal{C}_{1,1}^{E}=0$ and $\left(\beta_{0}+\beta_{1}\right)^{\prime} \mathcal{C}_{1,2}^{E}=0$. Thus, the vector $\beta_{0}+\beta_{1}$ is a CIV of order $((0,2),\{ \})$ and therefore $\beta(z)=\beta_{0}+\beta_{1} z$ is of "non-minimum" degree, one in this case rather than zero $\left(\beta_{0}+\beta_{1}\right)$. For a formal definition of minimum degree PCIVs see Bauer and Wagner (2003, Definition 4). In case $\mathcal{C}_{1,2}^{G} \neq 0$ there are PCIVs of degree one that are not simple transformations of static CIVs. Consider $\beta(z)=$ $\beta_{0}+\beta_{1} z=\gamma_{1}(1-z)+\gamma_{2}$ such that $\left\{\gamma_{1}^{\prime}\left(y_{t}-y_{t-1}\right)+\gamma_{2}^{\prime} y_{t}\right\}_{t \in \mathbb{Z}}$ is stationary. The integrated contribution to $\left\{\gamma_{1}^{\prime}\left(y_{t}-y_{t-1}\right)\right\}_{t \in \mathbb{Z}}$ is given by $\gamma_{1}^{\prime}(1-L)\left(\left\{\mathcal{C}_{1,1}^{E} x_{t, 1}^{E}\right\}_{t \in \mathbb{Z}}\right)=\left\{\gamma_{1}^{\prime} \mathcal{C}_{1,1}^{E} x_{t-1,2}^{G}+\gamma_{1}^{\prime} \mathcal{C}_{1,1}^{E} \mathcal{B}_{1,1} \varepsilon_{t-1}\right\}_{t \in \mathbb{Z}}$, with $\gamma_{1}^{\prime} \mathcal{C}_{1,1}^{E} \neq 0$. This term is eliminated by $\left\{\gamma_{2}^{\prime} \mathcal{C}_{1,2}^{G} x_{t, 2}^{G}\right\}_{t \in \mathbb{Z}}$ in $\left\{\gamma_{2}^{\prime} y_{t}\right\}_{t \in \mathbb{Z}}$, if $\gamma_{1}^{\prime} \mathcal{C}_{1,1}^{E}+\gamma_{2}^{\prime} \mathcal{C}_{1,2}^{G}=0$, which is only possible if $\mathcal{C}_{1,2}^{G} \neq 0$. Additionally, $\gamma_{2}^{\prime}\left[\mathcal{C}_{1,1}^{E} \mathcal{C}_{1,2}^{E}\right]=0$ needs to hold, such that there is no further integrated contribution to $\left\{\gamma_{2}^{\prime} y_{t}\right\}_{t \in \mathbb{Z}}$. Neither $\gamma_{1}$ nor $\gamma_{2}$ are CIVs since both violate the necessary conditions given in the definition of CIVs, which implies that $\beta(z)$ is indeed a "minimum degree" PCIV.

As was shown above, the unit root and cointegration properties of $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ depend on the sub-blocks of $\mathcal{C}_{u}$ and the eigenvalue structure of $\mathcal{A}_{u}$. We, therefore, define the more encompassing state space unit root structure containing information on the geometrical and algebraic multiplicities of the eigenvalues of $\mathcal{A}_{u}$ (cf. Bauer and Wagner 2012, Definition 2).

Definition 4. A unit root process $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ with a canonical state space representation as given in Theorem 1 has state space unit root structure

$$
\Omega_{S}:=\left(\left(\omega_{1}, d_{1}^{1}, \ldots, d_{h_{1}}^{1}\right), \ldots,\left(\omega_{l}, d_{1}^{l}, \ldots, d_{h_{l}}^{l}\right)\right)
$$

where $0 \leq d_{1}^{k} \leq d_{2}^{k} \leq \cdots \leq d_{h_{k}}^{k} \leq s$ for $k=1, \ldots, l$. For $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ with empty unit root structure $\Omega_{S}:=\{ \}$.

Remark 8. The state space unit root structure $\Omega_{S}$ contains information concerning the integration properties of the process $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$, since the integers $d_{j}^{k}, k=1, \ldots, l, j=1, \ldots, h_{k}$ describe (multiplied by two for $k$ such that $\left.0<\omega_{k}<\pi\right)$ the numbers of non-cointegrated stochastic trends or cycles of corresponding integration orders, compare again Remark 4. As such, $\Omega_{S}$ describes properties of the stochastic process $\left\{y_{t}\right\}_{t \in \mathbb{Z} —}$ and, therefore, the state space unit root structure $\Omega_{S}$ partitions unit root processes according to these (co-)integration properties. These (co-)integration properties, however, are invariant to a chosen canonical representation, or more generally invariant to whether a VARMA or state space representation is considered. For all minimal state representations of a unit root process $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ these indices-being related to the Jordan normal form-are invariant.

As mentioned in Section 2, Paruolo (1996, Definition 3) introduces integration indices at frequency zero as a triple of integers $\left(r_{0}, r_{1}, r_{2}\right)$. These correspond to the numbers of columns of the matrices $\beta, \beta_{1}, \beta_{2}$ in the error correction representation of $\mathrm{I}(2)$ VAR processes, see, e.g., Johansen (1997, sct. 3). Here, $r_{2}$ is the number of stochastic trends of order two, i.e., $r_{2}=d_{1}^{1}$. Furthermore, $r_{1}$ is the number of stochastic trends of order one that do not cointegrate with $\beta_{2}^{\prime} \Delta_{0}\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ and hence $r_{1}=d_{2}^{1}-d_{1}^{1}$. Therefore, the integration indices at frequency zero are in one-one correspondence with the state space unit root structure $\Omega_{S}=\left(\left(0, d_{1}^{1}, d_{2}^{1}\right)\right)$ for $\mathrm{I}(2)$ processes and the dimension $s=r_{0}+r_{1}+r_{2}$ of the process.

The canonical form given in Theorem 1 imposes p.u.t. structures on sub-blocks of the matrix $\mathcal{B}_{u}$. The occurrence of these blocks-related to $d_{j}^{k}>d_{j-1}^{k}$ —is determined by the state space unit root structure $\Omega_{S}$. The number of free entries in these p.u.t.-blocks, however, is not determined by $\Omega_{S}$. Consequently, we need structure indices $p \in \mathbb{N}_{0}^{n_{u}}$ indicating for each row the position of a potentially restricted positive element, as formalized below:

Definition 5 (Structure indices). For the block $\mathcal{B}_{u} \in \mathbb{C}^{n_{u} \times s}$ of the matrix $\mathcal{B}$ of a state space realization $(\mathcal{A}, \mathcal{B}, \mathcal{C})$ in canonical form, define the corresponding structure indices $p \in \mathbb{N}_{0}^{n_{u}}$ as

$$
p_{i}:= \begin{cases}0 & \text { if the } i \text {-th row of } \mathcal{B}_{u} \text { is not part of a p.u.t. block, } \\ j & \text { if the } i \text {-th row of } \mathcal{B}_{u} \text { is part of a p.u.t. block and its } j \text {-th entry is restricted to be positive. }\end{cases}
$$

Remark 9. Since sub-blocks of $\mathcal{B}_{u}$ corresponding to complex unit roots are of the form $\mathcal{B}_{k, \mathbb{C}}=\left[\mathcal{B}_{k^{\prime}}^{\prime} \overline{\mathcal{B}}_{k}^{\prime}\right]^{\prime}$, the entries restricted to be positive are located in the same columns and rows of both $\mathcal{B}_{k}$ and $\overline{\mathcal{B}}_{k}$. Thus, the structure indices $p_{i}$ of the corresponding rows are identical for $\mathcal{B}_{k}$ and $\overline{\mathcal{B}}_{k}$. Therefore, it would be possible to omit the parts of $p$ corresponding to the blocks $\overline{\mathcal{B}}_{k}$. It is, however, as will be seen in Definition 9, advantageous for the comparison of unit root structures and structure indices that $p$ is a vector with $n_{u}$ entries.

Example 4. Consider the following state space system:

$$
\begin{align*}
y_{t} & =\left[\begin{array}{lll}
\mathcal{C}_{1,1}^{E} & \mathcal{C}_{1,2}^{G} & \mathcal{C}_{1,2}^{E}
\end{array}\right] x_{t}+\varepsilon_{t} \quad y_{t}, \varepsilon_{t} \in \mathbb{R}^{2}, x_{t} \in \mathbb{R}^{3}, \mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{G}, \mathcal{C}_{1,2}^{E} \in \mathbb{R}^{2 \times 1}  \tag{12}\\
x_{t+1} & =\left[\begin{array}{lll}
1 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right] x_{t}+\left[\begin{array}{c}
\mathcal{B}_{1,1} \\
\mathcal{B}_{1,2,1} \\
\mathcal{B}_{1,2,2}
\end{array}\right] \varepsilon_{t,} \quad x_{0}=0, \quad \mathcal{B}_{1,1}, \mathcal{B}_{1,2,1}, \mathcal{B}_{1,2,2} \in \mathbb{R}^{1 \times 2}
\end{align*}
$$

In canonical form $\mathcal{B}_{1,2,1}$ and $\mathcal{B}_{1,2,2}$ are p.u.t. matrices and $\mathcal{B}_{1,1}$ is unrestricted. If, e.g., the second entry $b_{1,2,1,2}$ of $\mathcal{B}_{1,2,1}$ and the first entry $b_{1,2,2,1}$ of $\mathcal{B}_{1,2,2}$ are restricted to be positive, then

$$
\mathcal{B}=\left[\begin{array}{cc}
* & * \\
0 & b_{1,2,1,2} \\
b_{1,2,2,1} & *
\end{array}\right]
$$

where the symbol $*$ denotes unrestricted entries. In this case $p=[0,2,1]^{\prime}$.

For given state space unit root structure $\Omega_{S}$ the matrix $\mathcal{A}_{u}$ is fully determined. The parameterization of the set of feasible matrices $\mathcal{B}_{u}$ for given structure indices $p$ and of the set of stable subsystems $\left(\mathcal{A}_{\bullet}, \mathcal{B}_{\bullet}, \mathcal{C}_{\bullet}\right)$ for given Kronecker indices $\alpha_{\bullet}$ (cf. Hannan and Deistler 1988, chp. 2.) is straightforward, since the entries in these matrices are either unrestricted, restricted to zero or restricted to be positive. Matters are a bit more complicated for $\mathcal{C}_{u}$. One possibility to parameterize the set of possible matrices $\mathcal{C}_{u}$ for a given state space unit root structure $\Omega_{S}$ is to use real and complex valued Givens rotations (cf. Golub and van Loan 1996, chp. 5.1).

Definition 6 (Real Givens rotation). The real Givens rotation $R_{q, i, j}(\theta) \in \mathbb{R}^{q \times q}, \theta \in[0,2 \pi)$ is defined as

$$
R_{q, i, j}(\theta):=\left[\begin{array}{ccccc}
I_{i-1} & & & & 0 \\
& \cos (\theta) & 0 & \sin (\theta) & \\
& 0 & I_{j-1-i} & 0 & \\
& -\sin (\theta) & 0 & \cos (\theta) & \\
0 & & & & I_{q-j}
\end{array}\right]
$$

Remark 10. Givens rotations allow transforming any vector $v=\left[v_{1}, v_{2}, \ldots, v_{q}\right]^{\prime} \in \mathbb{R}^{q}$ into a vector of the form $\left[\tilde{v}_{1}, 0, \ldots, 0\right]^{\prime}$ with $\tilde{v}_{1} \geq 0$. This is achieved by the following algorithm:

1. $\quad$ Set $j=1, v_{1}^{(1)}=v_{1}$ and $v^{(1)}=v$.
2. Represent $\left[v_{1}^{(j)}, v_{q-j+1}\right]^{\prime}$ using polar coordinates as $\left[v_{1}^{(j)}, v_{q-j+1}\right]^{\prime}=\left[r_{j} \cos \left(\theta_{q-j}\right), r_{j} \sin \left(\theta_{q-j}\right)\right]^{\prime}$, with $r_{j} \geq 0$ and $\theta_{q-j} \in[0,2 \pi)$. If $r_{j}=0$, set $\theta_{q-j}=0$ (cf. Otto 2011, chp. 1.5.3, $p$. 39). Then $R_{2,1,2}\left(\theta_{q-j}\right)\left[v_{1}^{(j)}, v_{q-j+1}\right]^{\prime}=\left[v_{1}^{(j+1)}, 0\right]^{\prime}$ such that $v^{(j+1)}=R_{q, 1, q-j+1}\left(\theta_{q-j}\right) v^{(j)}=$ $\left[v_{1}^{(j+1)}, v_{2}, \ldots, v_{q-j}, 0, \ldots, 0\right]^{\prime}$, with $v_{1}^{(j+1)} \geq 0$.
3. If $j=q-1$, stop. Else increment $j$ by one $(j \rightarrow j+1)$ and continue at step 2 .

This algorithm determines a unique vector $\theta=\left[\theta_{1}, \ldots, \theta_{q-1}\right]^{\prime}$ for every vector $v \in \mathbb{R}^{q}$.
Remark 11. The determinant of real Givens rotations is equal to one, i.e., $\operatorname{det}\left(R_{s, i, j}(\theta)\right)=1$ for all $s, i, j \in \mathbb{N}$ and all $\theta \in[0,2 \pi)$. Thus, it is not possible to factorize an orthonormal matrix $Q$ with $\operatorname{det}(Q)=-1$ into a product of Givens rotations. This obvious fact has implications for the parameterization of $\mathcal{C}$-matrices as is detailed below.

Definition 7 (Complex Givens rotation). The complex Givens rotation $Q_{q, i, j}(\varphi) \in \mathbb{C}^{q \times q}, \varphi:=$ $\left[\varphi_{1}, \varphi_{2}\right]^{\prime} \in \Theta_{\mathbb{C}}:=[0, \pi / 2] \times[0,2 \pi)$, is defined as

$$
Q_{q, i, j}(\varphi):=\left[\begin{array}{ccccc}
I_{i-1} & & & & 0 \\
& \cos \left(\varphi_{1}\right) & 0 & \sin \left(\varphi_{1}\right) e^{i \varphi_{2}} & \\
& 0 & I_{j-1-i} & 0 & \\
& -\sin \left(\varphi_{1}\right) e^{-i \varphi_{2}} & 0 & \cos \left(\varphi_{1}\right) & \\
0 & & & & I_{q-j}
\end{array}\right]
$$

Remark 12. Complex Givens rotations allow transforming any vector $v=\left[v_{1}, v_{2}, \ldots, v_{q}\right]^{\prime} \in \mathbb{C}^{q}$ into a vector of the form $\left[\tilde{v}_{1}, 0, \ldots, 0\right]^{\prime}$ with $\tilde{v}_{1} \in \mathbb{C}$. This is achieved by the following algorithm:

1. Set $j=1, v_{1}^{(1)}=v_{1}$ and $v^{(1)}=v$.
2. Represent $\left[v_{1}^{(j)}, v_{q-j+1}\right]^{\prime}$ using polar coordinates as $\left[v_{1}^{(j)}, v_{q-j+1}\right]^{\prime}=\left[a_{j} e^{i \varphi_{a, j}}, b_{j} e^{i \varphi_{b, j}}\right]^{\prime}$, with $a_{j}, b_{j} \geq 0$ and $\varphi_{a, j}, \varphi_{b, j} \in[0,2 \pi)$. If $v_{1}^{(j)}=0$, set $\varphi_{a, j}=0$ and if $v_{q-j+1}=0$, set $\varphi_{b, j}=0$ (cf. Otto 2011, chp. 8.1.3, p. 222).
3. Set

$$
\begin{aligned}
& \varphi_{q-j, 1}= \begin{cases}\tan ^{-1}\left(\frac{b_{j}}{a_{j}}\right) & \text { if } a_{j}>0 \\
\pi / 2 & \text { if } a_{j}=0, b_{j}>0 \\
0 & \text { if } a_{j}=0, b_{j}=0\end{cases} \\
& \varphi_{q-j, 2}=\varphi_{a, j}-\varphi_{b, j} \bmod 2 \pi
\end{aligned}
$$

Then $Q_{2,1,2}\left(\varphi_{q-j}\right)\left[v_{1}^{(j)}, v_{q-j+1}\right]^{\prime}=\left[v_{1}^{(j+1)}, 0\right]^{\prime}$ such that $v^{(j+1)}=Q_{q, 1, q-j+1}\left(\theta_{q-1}\right) v^{(j)}=$ $\left[v_{1}^{(j+1)}, v_{2}, \ldots, v_{q-j}, 0\right]^{\prime}$, with $v_{1}^{(j+1)} \in \mathbb{C}$.
4. If $j=q-1$, stop. Else increment $j$ by one $(j \rightarrow j+1)$ and continue at step 2 .

This algorithm determines a unique vector $\varphi=\left[\varphi_{1,1}, \varphi_{1,2}, \ldots, \varphi_{q-1,2}\right]^{\prime}$ for every vector $v \in \mathbb{C}^{q}$.
To set the stage for the general case, we start the discussion of the parameterization of the set of matrices $(\mathcal{A}, \mathcal{B}, \mathcal{C})$ in canonical form with the $\operatorname{MFI}(1)$ and $\mathrm{I}(2)$ cases. These two cases display all ingredients required later for the general case. The $\mathrm{MFI}(1)$ case illustrates the usage of either real or complex Givens rotations, depending on whether the considered $\mathcal{C}$-block corresponds to a real or complex unit root. The $I(2)$ case highlights recursive orthogonality constraints on the parameters of the $\mathcal{C}$-block, which are related to the polynomial cointegration properties (cf. Example 3).

### 3.1. The Parameterization in the MFI(1) Case

The state space unit root structure of an $\operatorname{MFI}(1)$ process is given by $\Omega_{S}=\left(\left(\omega_{1}, d_{1}^{1}\right), \ldots,\left(\omega_{l}, d_{1}^{l}\right)\right)$. For the corresponding state space system $(\mathcal{A}, \mathcal{B}, \mathcal{C})$ in canonical form, the sub-blocks of $\mathcal{A}_{u}$ are equal to $J_{k}=\overline{z_{k}} I_{d_{1}^{k}}$, the sub-blocks $\mathcal{B}_{k}$ of $\mathcal{B}_{u}$ are p.u.t. and $\mathcal{C}_{k}^{\prime} \mathcal{C}_{k}=I_{d_{1}^{k}}$, for $k=1, \ldots, l$.

Starting with the sub-blocks of $\mathcal{C}_{u}$, it is convenient to separate the discussion of the parameterization of $\mathcal{C}_{u}$-blocks into the real case, where $\omega_{k} \in\{0, \pi\}$ and $\mathcal{C}_{k} \in \mathbb{R}^{s \times d_{1}^{k}}$, and the complex case with $0<\omega_{k}<\pi$ and $\mathcal{C}_{k} \in \mathbb{C}^{s \times d_{1}^{k}}$. For the case of real unit roots the two cases $d_{1}^{k}<s$ and $d_{1}^{k}=s$ have to be distinguished. For brevity of notation refer to the considered real block simply as $\mathcal{C} \in \mathbb{R}^{s \times d}$. Using this notation, the set of matrices to be parameterized is

$$
O_{s, d}:=\left\{C \in \mathbb{R}^{s \times d} \mid C^{\prime} C=I_{d}\right\} .
$$

The parameterization of $O_{s, d}$ is based on the combination of real Givens rotations, as given in Definition 6, that allow transforming every matrix in $O_{s, d}$ to the form $\left[I_{d}, 0_{(s-d) \times d}^{\prime}\right]^{\prime}$ for $d<s$. For $d=s$, Givens rotations allow transforming every matrix $\mathcal{C} \in O_{s, s}$ either to $I_{s}$ or $I_{s}^{-}:=\operatorname{diag}\left(I_{s-1},-1\right)$, since, compare Remark 11, for the transformed matrix $\tilde{\mathcal{C}}^{(s)}$ it holds that $\operatorname{det}(\mathcal{C})=\operatorname{det}\left(\tilde{\mathcal{C}}^{(s)}\right) \in\{-1,1\}$. This is achieved with the following algorithm:

1. Set $j=1$ and $\mathcal{C}^{(1)}=\mathcal{C}$.
2. Transform the entries $\left[c_{j, j}, \ldots, c_{j, d}\right]$ in the $j$-th row of $\mathcal{C}(j)$, to $\left[\tilde{c}_{j, j}, 0, \ldots, 0\right], \tilde{c}_{j, j} \geq 0$. Since this is a row vector, this is achieved by right-multiplication of $\mathcal{C}^{(j)}$ with transposed Givens rotations and the required parameters are obtained via the algorithm described in Remark 10. The first $j-1$ entries of the $j$-th row remain unchanged. Denote the transformed matrix by $\mathcal{C}^{(j+1)}$.
3. If $j=d-1$ stop. Else increment $j$ by one $(j \rightarrow j+1)$ and continue at step 2 .
4. Collect all parameters used for the Givens rotations in steps 1 to 3 in a parameter vector $\boldsymbol{\theta}_{R}$. Steps 1-3 correspond to a $Q R$ decomposition of $\mathcal{C}^{\prime}=Q \tilde{\mathcal{C}}^{\prime}$, with an orthonormal matrix $Q$ given by the product of the Givens rotations. Please note that the first $j-1$ entries of the $j$-th column of $\tilde{\mathcal{C}}=\mathcal{C}^{(d)}$ are equal to zero by construction.
5. Set $j=0$ and $\tilde{\mathcal{C}}^{(0)}=\tilde{\mathcal{C}}$.
6. Collect the entries in column $d-j$ of $\tilde{\mathcal{C}}^{(j)}$ which have not been transformed to zero by previous transformations into the vector $\left[c_{d-j, d-j}, c_{d+1, d-j}, \ldots, c_{s, d-j}\right]^{\prime}$. Using the algorithm described in Remark 10 transform this vector to $\left[\tilde{c}_{d-j, d-j}, 0, \ldots, 0\right]^{\prime}$ by left-multiplication of $\tilde{\mathcal{C}}^{(j)}$ with Givens rotations. Since Givens rotations are orthonormal, the transformed matrix $\tilde{\mathcal{C}}^{(j+1)}$ is still orthonormal implying for its entries $\tilde{c}_{d-j, d-j}=1$ and $\tilde{c}_{i, d-j}=0$ for all $i<d-j$. An exception occurs if $d=s$. In this case $c_{d-j, d-j} \in\{-1,1\}$ and no Givens rotations are defined.
7. If $j=d-1$ stop. Else increment $j$ by one $(j \rightarrow j+1)$ and continue at step 6 .
8. Collect all parameters used for the Givens rotations in steps 5 to 7 in a parameter vector $\boldsymbol{\theta}_{L}$.

The parameter vector $\boldsymbol{\theta}=\left[\boldsymbol{\theta}_{L}^{\prime}, \boldsymbol{\theta}_{R}^{\prime}\right]^{\prime}$, contains the angles of the employed Givens rotations and provides one way of parameterizing $O_{s, d}$. The following Lemma 1 demonstrates the usefulness of this parameterization.

Lemma 1 (Properties of the parameterization of $O_{s, d}$ ). Define for $d \leq s$ a mapping $\boldsymbol{\theta} \rightarrow C_{O}(\boldsymbol{\theta})$ from $\Theta_{O}^{\mathbb{R}}:=[0,2 \pi)^{d(s-d)} \times[0,2 \pi)^{d(d-1) / 2} \rightarrow O_{s, d} b y$

$$
\begin{aligned}
C_{O}(\boldsymbol{\theta}) & :=\left[\prod_{i=1}^{d} \prod_{j=1}^{s-d} R_{s, i, d+j}\left(\theta_{L,(s-d)(i-1)+j}\right)\right]^{\prime}\left[\begin{array}{c}
I_{d} \\
0_{(s-d) \times d}
\end{array}\right]\left[\prod_{i=1}^{d-1} \prod_{j=1}^{i} R_{d, d-i, d-i+j}\left(\theta_{R, i(i-1) / 2+j}\right)\right] \\
& :=R_{L}\left(\boldsymbol{\theta}_{L}\right)^{\prime}\left[\begin{array}{c}
I_{d} \\
0_{(s-d) \times d}
\end{array}\right] R_{R}\left(\boldsymbol{\theta}_{R}\right),
\end{aligned}
$$

with $\boldsymbol{\theta}:=\left[\boldsymbol{\theta}_{L}^{\prime}, \boldsymbol{\theta}_{R}^{\prime}\right]^{\prime}$, where $\boldsymbol{\theta}_{L}:=\left[\theta_{L, 1}, \ldots, \theta_{L, d(s-d)}\right]^{\prime}$ and $\boldsymbol{\theta}_{R}:=\left[\theta_{R, 1}, \ldots, \theta_{R, d(d-1) / 2}\right]^{\prime}$. The following properties hold:
(i) $O_{s, d}$ is closed and bounded.
(ii) The mapping $C_{O}(\cdot)$ is infinitely often differentiable.

For $d<s$, it holds that
(iii) For every $\mathcal{C} \in O_{s, d}$ there exists a vector $\boldsymbol{\theta} \in \Theta_{O}^{\mathbb{R}}$ such that

$$
\mathcal{C}=C_{O}(\boldsymbol{\theta})=R_{L}\left(\boldsymbol{\theta}_{L}\right)^{\prime}\left[\begin{array}{c}
I_{d} \\
0_{(s-d) \times d}
\end{array}\right] R_{R}\left(\boldsymbol{\theta}_{R}\right) .
$$

The algorithm discussed above defines the inverse mapping $C_{O}^{-1}: O_{s, d} \rightarrow \Theta_{O}^{\mathbb{R}}$.
(iv) The inverse mapping $C_{O}^{-1}(\cdot)$-the parameterization of $O_{s, d}$-is infinitely often differentiable on the pre-image of the interior of $\Theta_{O}^{\mathbb{R}}$. This is an open and dense subset of $O_{s, d}$.

For $d=s$, it holds that
(v) $O_{s, s}$ is a disconnected space in $\mathbb{R}^{s \times s}$ with two disjoint non-empty closed subsets $O_{s, s}^{+}:=\left\{C \in \mathbb{R}^{s \times s} \mid C^{\prime} C=\right.$ $\left.I_{s}, \operatorname{det}(C)=1\right\}$ and $O_{s, s}^{-}:=\left\{C \in \mathbb{R}^{s \times s} \mid C^{\prime} C=I_{s}, \operatorname{det}(C)=-1\right\}$.
(vi) For every $\mathcal{C} \in O_{s, s}^{+}$there exists a vector $\boldsymbol{\theta} \in \Theta_{O}^{\mathbb{R}}$ such that

$$
\mathcal{C}=C_{O}(\boldsymbol{\theta})=R_{L}\left(\boldsymbol{\theta}_{L}\right)^{\prime}\left[I_{d}\right] R_{R}\left(\boldsymbol{\theta}_{R}\right)=R_{R}\left(\boldsymbol{\theta}_{R}\right)
$$

In this case, steps 1-4 of the algorithm discussed above define the inverse mapping $C_{O}^{-1}: O_{s, S}^{+} \rightarrow \Theta_{O}^{\mathbb{R}}$.
(vii) Define $v:=[\pi, \ldots, \pi]^{\prime} \in \mathbb{R}^{s(s-1) / 2}$. Then a parameterization of $O_{s, s}$ is given by

$$
C_{O}^{ \pm}(C)= \begin{cases}v+C_{O}^{-1}(C) & \text { if } C \in O_{s, s}^{+} \\ -\left(v+C_{O}^{-1}\left(C I_{s}^{-}\right)\right) & \text {if } C \in O_{s, s}^{-}\end{cases}
$$

The parameterization is infinitely often differentiable with infinitely often differentiable inverse on an open and dense subset of $O_{s, s}$.

Remark 13. The following arguments illustrate why $C_{O}^{-1}$ is not continuous on the pre-image of the boundary of $\Theta_{O}^{\mathbb{R}}$ : Consider the unit sphere $O_{3,1}=\left\{C \in \mathbb{R}^{3} \mid C^{\prime} C=\|C\|_{2}=1\right\}$. One way to parameterize the unit sphere is to use degrees of longitude and latitude. Two types of discontinuities occur: After fixing the location of the zero degree of longitude, i.e., the prime meridian, its anti-meridian is described by both $180^{\circ} \mathrm{W}$ and $180^{\circ} \mathrm{E}$. Using the half-open interval $[0,2 \pi)$ in our parametrization causes a similar discontinuity. Second, the degree of longitude is irrelevant at the north pole. As seen in Remark 10, with our parameterization a similar issue occurs when the first two entries of $C$ to be compared are both equal to zero. In this case the parameter of the Givens rotation is set to zero, although every $\theta$ will produce the same result. Both discontinuities clearly occur on a thin subset of $O_{s, d}$.

As in the parametrization of the VAR I(1)-case in the VECM framework, where the restriction $\beta=$ $\left[I_{s-d}, \beta^{*}\right]^{\prime}$ can only be imposed when the upper $(s-d) \times(s-d)$ block of the true $\beta_{0}$ of the DGP is of full rank (cf. Johansen 1995, chp. 5.2), the set where the discontinuities occur can effectively be changed by a permutation of the components of the observed time series. This corresponds to redefining the locations of the prime meridian and the poles.

Remark 14. Please note that the parameterization partitions the parameter vector $\boldsymbol{\theta}$ into two parts $\boldsymbol{\theta}_{L} \in$ $[0,2 \pi)^{d(s-d)}$ and $\boldsymbol{\theta}_{R} \in[0,2 \pi)^{(d-1) d / 2}$. Since changing the parameter values in $\boldsymbol{\theta}_{R}$ does not change the column space of $C_{O}(\boldsymbol{\theta})$, which, as seen above, determines the cointegrating vectors, $\boldsymbol{\theta}_{L}$ fully characterizes the (static) cointegrating space. Please note that the dimension of $\boldsymbol{\theta}_{L}$ is $d(s-d)$ and thus coincides with the number of free parameters in $\beta$ in the VECM framework (cf. Johansen 1995, chp. 5.2).

Example 5. Consider the matrix

$$
C=\left[\begin{array}{cc}
0 & \frac{1}{\sqrt{2}} \\
\frac{-1}{\sqrt{2}} & \frac{1}{2} \\
\frac{1}{\sqrt{2}} & \frac{1}{2}
\end{array}\right]
$$

with $d=2$ and $s=3$. As discussed, the static cointegrating space is characterized by the left kernel of this matrix. The left kernel of a matrix in $\mathbb{R}^{3 \times 2}$ with full rank two is given by a one-dimensional space, with the corresponding basis vector parameterized, when normalized to length one, by two free parameters. Thus, for the characterization of the static cointegrating space two parameters are required, which exactly coincides with the dimension of $\boldsymbol{\theta}_{L}$ given in Remark 14. The parameters in $\boldsymbol{\theta}_{R}$ correspond to the choice of a basis of the image of $C$. Having fixed the two-dimensional subspace through $\boldsymbol{\theta}_{L}$, only one free parameter for the choice of an orthonormal basis remains, which again coincides with the dimension given in Remark 14. To obtain the parameter vector, the starting point is a $Q R$ decomposition of $C^{\prime}=R_{R}\left(\boldsymbol{\theta}_{R}\right) \tilde{C}^{\prime}$. In this example $R_{R}\left(\boldsymbol{\theta}_{R}\right)=R_{2,1,2}\left(\theta_{R, 1}\right)$, with $\theta_{R, 1}$ to be determined. To find $\theta_{R, 1}$, solve $\left[\begin{array}{cc}0 & \frac{1}{\sqrt{2}}\end{array}\right] R_{2,1,2}\left(\theta_{R, 1}\right)^{\prime}=\left[\begin{array}{ll}r & 0\end{array}\right]$ for $r \geq 0$ and $\theta_{R, 1} \in[0,2 \pi)$. In other words, find $r \geq 0$ and $\theta_{R, 1} \in[0,2 \pi)$ such that $\left[\begin{array}{cc}0 & \frac{1}{\sqrt{2}}\end{array}\right]=r\left[\cos \left(\theta_{R, 1}\right) \sin \left(\theta_{R, 1}\right)\right]$, which leads to $r=\frac{1}{\sqrt{2}}, \theta_{R, 1}=\frac{\pi}{2}$. Thus, the orthonormal matrix $R_{R}\left(\boldsymbol{\theta}_{R}\right)$ is equal to $R_{2,1,2}\left(\frac{\pi}{2}\right)$ and the transpose of the upper triangular matrix $\tilde{C}^{\prime}$ is equal to:

$$
\tilde{C}=\tilde{C}^{(0)}=C \cdot R_{2,1,2}\left(\frac{\pi}{2}\right)^{\prime}=\left[\begin{array}{cc}
0 & \frac{1}{\sqrt{2}} \\
\frac{-1}{\sqrt{2}} & \frac{1}{2} \\
\frac{1}{\sqrt{2}} & \frac{1}{2}
\end{array}\right]\left[\begin{array}{cc}
0 & -1 \\
1 & 0
\end{array}\right]=\left[\begin{array}{cc}
\frac{1}{\sqrt{2}} & 0 \\
\frac{1}{2} & \frac{1}{\sqrt{2}} \\
\frac{1}{2} & -\frac{1}{\sqrt{2}}
\end{array}\right] .
$$

Second, transform the entries in the lower $1 \times 2$-sub-block of $\tilde{C}^{(0)}$ to zero, starting with the last column. For this find $\theta_{L, 2} \in[0,2 \pi)$ such that $R_{3,2,3}\left(\theta_{L, 2}\right)\left[\begin{array}{lll}0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\end{array}\right]^{\prime}=\left[\begin{array}{lll}0 & 1 & 0\end{array}\right]^{\prime}$, i.e., $\left[\begin{array}{ll}\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\end{array}\right]^{\prime}=$ $r\left[\cos \left(\theta_{L, 2}\right) \sin \left(\theta_{L, 2}\right)\right]$. This yields $r=1, \theta_{L, 2}=\frac{7 \pi}{4}$. Next compute $\tilde{C}^{(1)}=R_{3,2,3}\left(\frac{7 \pi}{4}\right) \tilde{C}^{(0)}$ :

$$
\tilde{C}^{(1)}=R_{3,2,3}\left(\frac{7 \pi}{4}\right) \cdot C \cdot R_{2,1,2}\left(\frac{\pi}{2}\right)^{\prime}=\left[\begin{array}{ccc}
1 & 0 & 0 \\
0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\
0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}
\end{array}\right]\left[\begin{array}{cc}
0 & \frac{1}{\sqrt{2}} \\
\frac{-1}{\sqrt{2}} & \frac{1}{2} \\
\frac{1}{\sqrt{2}} & \frac{1}{2}
\end{array}\right]\left[\begin{array}{cc}
0 & -1 \\
1 & 0
\end{array}\right]=\left[\begin{array}{cc}
\frac{1}{\sqrt{2}} & 0 \\
0 & 1 \\
\frac{1}{\sqrt{2}} & 0
\end{array}\right] .
$$

In the final step find $\theta_{L, 1} \in[0,2 \pi)$ such that $R_{3,1,3}\left(\theta_{L, 1}\right)\left[\begin{array}{lll}\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}}\end{array}\right]^{\prime}=\left[\begin{array}{lll}1 & 0 & 0\end{array}\right]^{\prime}$, i.e., $\left[\begin{array}{ll}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{array}\right]^{\prime}=r\left[\cos \left(\theta_{L, 1}\right) \quad \sin \left(\theta_{L, 1}\right)\right]$. The solution is $r=1, \theta_{L, 1}=\frac{\pi}{4}$. Combining the transformations leads to

$$
\begin{gathered}
R_{3,1,3}\left(\frac{\pi}{4}\right) \cdot R_{3,2,3}\left(\frac{7 \pi}{4}\right) \cdot C \cdot R_{2,1,2}\left(\frac{\pi}{2}\right)^{\prime}= \\
{\left[\begin{array}{ccc}
\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\
0 & 1 & 0 \\
\frac{-1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}}
\end{array}\right]\left[\begin{array}{ccc}
1 & 0 & 0 \\
0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\
0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}
\end{array}\right]\left[\begin{array}{cc}
0 & \frac{1}{\sqrt{2}} \\
\frac{-1}{\sqrt{2}} & \frac{1}{2} \\
\frac{1}{\sqrt{2}} & \frac{1}{2}
\end{array}\right]\left[\begin{array}{cc}
0 & -1 \\
1 & 0
\end{array}\right]=\left[\begin{array}{cc}
1 & 0 \\
0 & 1 \\
0 & 0
\end{array}\right] .}
\end{gathered}
$$

The parameter vector for this matrix is therefore $\boldsymbol{\theta}=\left[\boldsymbol{\theta}_{L}^{\prime}, \boldsymbol{\theta}_{R}^{\prime}\right]^{\prime}=\left[\left[\frac{\pi}{4}, \frac{7 \pi}{4}\right],\left[\frac{\pi}{2}\right]\right]^{\prime}$ with $\boldsymbol{\theta}=C_{O}^{-1}(C)$.
In case of complex unit roots, referring for brevity again to the considered block $\mathcal{C}_{k}$ simply as $\mathcal{C} \in \mathbb{C}^{s \times d}$, the set of matrices to be parameterized is

$$
U_{s, d}:=\left\{C \in \mathbb{C}^{s \times d} \mid C^{\prime} C=I_{d}\right\} .
$$

The parameterization of this set is based on the combination of complex Givens rotations, as given in Definition 7, which can be used to transform every matrix in $U_{s, d}$ to the form $\left[D_{d}, 0_{(s-d) \times d}^{\prime}\right]^{\prime}$ with a diagonal matrix $D_{d}$ whose diagonal elements are of unit modulus. This transformation is achieved with the following algorithm:

1. Set $j=1$ and $\mathcal{C}^{(1)}=\mathcal{C}$.
2. Transform the entries $\left[c_{j, j}, \ldots, c_{j, d}\right]$ in the $j$-th row of $\mathcal{C}^{(j)}$, to $\left[\tilde{c}_{j, j}, 0, \ldots, 0\right]$. Since this is a row vector, this is achieved by right-multiplication of $\mathcal{C}$ with transposed Givens rotations and the required parameters are obtained via the algorithm described in Remark 12. The first $j-1$ entries of the $j$-th row remain unchanged. Denote the transformed matrix by $\mathcal{C}^{(j+1)}$.
3. If $j=d-1$ stop. Else increment $j$ by one $(j \rightarrow j+1)$ and continue at step 2 .
4. Collect all parameters used for the Givens rotations in steps 1 to 3 in a parameter vector $\varphi_{R}$. Step 1-3 corresponds to a $Q R$ decomposition of $\mathcal{C}^{\prime}=Q \tilde{\mathcal{C}}^{\prime}$, with a unitary matrix $Q$ given by the product of the Givens rotations. Please note that the first $j-1$ entries of the $j$-th column of $\tilde{\mathcal{C}}=\mathcal{C}^{(d)}$ are equal to zero by construction.
5. Set $j=0$ and $\tilde{\mathcal{C}}^{(0)}=\tilde{\mathcal{C}}$.
6. Collect the entries in column $d-j$ of $\tilde{\mathcal{C}}^{(j)}$ which have not been transformed to zero by previous transformations into the vector $\left[c_{d-j, d-j}, c_{d+1, d-j}, \ldots, c_{s, d-j}\right]^{\prime}$. Using the algorithm described in Remark 12 transform this vector to $\left[\tilde{\mathcal{c}}_{d-j, d-j}, 0, \ldots, 0\right]^{\prime}$ by left-multiplication of $\tilde{\mathcal{C}}^{(j)}$ with Givens rotations. Since Givens rotations are unitary, the transformed matrix $\tilde{\mathcal{C}}^{(j+1)}$ is still unitary implying for its entries $\left|\tilde{c}_{d-j, d-j}\right|=1$ and $\tilde{c}_{i, d-j}=0$ for all $i<d-j$. An exception occurs if $d=s$. In this case $\left|c_{d-j, d-j}\right|=1$ and no Givens rotations are defined.
7. If $j=d-1$ stop. Else increment $j$ by one $(j \rightarrow j+1)$ and continue at step 6 .
8. Collect all parameters used for the Givens rotations in steps 5 to 7 in a parameter vector $\boldsymbol{\varphi}_{L}$.
9. Transform the diagonal entries of the transformed matrix $\tilde{\mathcal{C}}^{(d)}=\left[D_{d}, 0_{(s-d) \times d}^{\prime}\right]^{\prime}$ into polar coordinates and collect the angles in a parameter vector $\varphi_{D}$.

The following lemma demonstrates the usefulness of this parameterization.
Lemma 2 (Properties of the parametrization of $U_{s, d}$. Define for $d \leq s$ a mapping $\boldsymbol{\varphi} \rightarrow C_{U}(\boldsymbol{\varphi})$ from $\Theta_{U}^{\mathbb{C}}:=\Theta_{\mathbb{C}}^{d(s-d)} \times \Theta_{\mathbb{C}}^{(d-1) d / 2} \times[0,2 \pi)^{d} \rightarrow U_{s, d} b y$

$$
\begin{aligned}
C_{U}(\boldsymbol{\varphi}) & :=\left[\prod_{i=1}^{d} \prod_{j=1}^{s-d} Q_{s, i, d+j}\left(\varphi_{L,(s-d)(i-1)+j}\right)\right]^{\prime}\left[\begin{array}{c}
D_{d}\left(\boldsymbol{\varphi}_{D}\right) \\
0_{(s-d) \times d}
\end{array}\right]\left[\prod_{i=1}^{d-1} \prod_{j=1}^{i} Q_{d, d-i, d-i+j}\left(\varphi_{R, i(i-1) / 2+j}\right)\right] \\
& :=Q_{L}\left(\boldsymbol{\varphi}_{L}\right)^{\prime}\left[\begin{array}{c}
D_{d}\left(\boldsymbol{\varphi}_{D}\right) \\
0_{(s-d) \times d}
\end{array}\right] Q_{R}\left(\boldsymbol{\varphi}_{R}\right),
\end{aligned}
$$

with $\boldsymbol{\varphi}:=\left[\boldsymbol{\varphi}_{L}^{\prime}, \boldsymbol{\varphi}_{R}^{\prime}, \boldsymbol{\varphi}_{D}^{\prime}\right]^{\prime}$, where $\boldsymbol{\varphi}_{L}=\left[\varphi_{L, 1}, \ldots, \varphi_{L, d(s-d)}\right]^{\prime}, \boldsymbol{\varphi}_{R}:=\left[\varphi_{R, 1}, \ldots, \varphi_{R, d(d-1) / 2}\right]^{\prime}$ and $\boldsymbol{\varphi}_{D}:=$ $\left[\varphi_{D, 1}, \ldots, \varphi_{D, d}\right]^{\prime}$ and where $D_{d}\left(\varphi_{D}\right)=\operatorname{diag}\left(e^{i \varphi_{D, 1}}, \ldots, e^{i \varphi_{D, d}}\right)$. The following properties hold:
(i) $U_{s, d}$ is closed and bounded.
(ii) The mapping $C_{U}(\boldsymbol{\varphi})$ is infinitely often differentiable.
(iii) For every $\mathcal{C} \in U_{s, d}$ a vector $\boldsymbol{\varphi} \in \Theta_{U}^{\mathbb{C}}$ exists such that

$$
\mathcal{C}=C_{U}(\boldsymbol{\varphi})=Q_{L}\left(\boldsymbol{\varphi}_{L}\right)^{\prime}\left[\begin{array}{c}
D_{d}\left(\boldsymbol{\varphi}_{D}\right) \\
0_{(s-d) \times d}
\end{array}\right] Q_{R}\left(\boldsymbol{\varphi}_{R}\right) .
$$

The algorithm discussed above defines the inverse mapping $C_{U}^{-1}: U_{s, d} \rightarrow \Theta_{U}^{\mathbb{R}}$.
(iv) The inverse mapping $C_{U}^{-1}(\cdot)$ —the parameterization of $U_{s, d}$ —is infinitely often differentiable on an open and dense subset of $U_{s, d}$.

Remark 15. Note the partitioning of the parameter vector $\boldsymbol{\varphi}$ into the parts $\boldsymbol{\varphi}_{L}, \boldsymbol{\varphi}_{D}$ and $\boldsymbol{\varphi}_{R}$. The component $\boldsymbol{\varphi}_{L}$ fully characterizes the column space of $C_{U}(\boldsymbol{\varphi})$, i.e., $\varphi_{L}$ determines the cointegrating spaces.

Example 6. Consider the matrix

$$
C=\left[\begin{array}{cc}
\frac{1-i}{2} & \frac{1-i}{2} \\
\frac{1+i}{2} & \frac{-1-i}{2} \\
0 & 0
\end{array}\right]
$$

The starting point is again a $Q R$ decomposition of $C^{\prime}=Q_{R}\left(\boldsymbol{\varphi}_{R}\right) \tilde{C}^{\prime}=Q_{2,1,2}\left(\varphi_{R, 1}\right) \tilde{C}^{\prime}$. To find a complex Givens rotation such that $\left[\begin{array}{cc}\frac{1-i}{2} & \frac{1-i}{2}\end{array}\right] Q_{2,1,2}\left(\varphi_{R, 1}\right)^{\prime}=\left[\begin{array}{cc}r e^{i \varphi_{a}} & 0\end{array}\right]$ with $r>0$, transform the entries of $\left[\begin{array}{cc}\frac{1-i}{2} & \frac{1-i}{2}\end{array}\right]^{\prime}$ into polar coordinates. The equation $\left[\begin{array}{cc}\frac{1-i}{2} & \frac{1-i}{2}\end{array}\right]^{\prime}=\left[\begin{array}{ll}a e^{i} \varphi_{a} & b e^{i \varphi_{b}}\end{array}\right]^{\prime}$ has the solutions $a=b=\frac{1}{\sqrt{2}}$ and $\varphi_{a}=\varphi_{b}=\frac{7 \pi}{4}$. Using the results of Remark 12, the parameters of the Givens rotation are $\varphi_{R, 1,1}=\tan ^{-1}\left(\frac{b}{a}\right)=\frac{\pi}{4}$ and $\varphi_{R, 1,2}=\varphi_{a}-\varphi_{b}=0$. Right-multiplication of $C$ with $Q_{2,1,2}\left(\left[\frac{\pi}{4}, 0\right]\right)^{\prime}$ leads to

$$
\tilde{C}=C Q_{2,1,2}\left(\left[\frac{\pi}{4}, 0\right]\right)^{\prime}=C\left[\begin{array}{cc}
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}}
\end{array}\right]^{\prime}=\left[\begin{array}{cc}
\frac{1-i}{\sqrt{2}} & 0 \\
0 & \frac{-1-i}{\sqrt{2}} \\
0 & 0
\end{array}\right]=\left[\begin{array}{c}
D_{2}\left(\boldsymbol{\varphi}_{D}\right) \\
0_{1 \times 2}
\end{array}\right] .
$$

Since the entries in the lower $1 \times 2$-sub-block of $\tilde{C}$ are already equal to zero, the remaining complex Givens rotations are $Q_{3,2,3}([0,0])=Q_{3,1,3}([0,0])=I_{3}$. Finally, the parameter values corresponding to the diagonal matrix $D_{2}\left(\varphi_{D}\right)=\operatorname{diag}\left(e^{i \varphi_{D, 1}}, e^{i \varphi_{D, 2}}\right)=\operatorname{diag}\left(\frac{1-i}{\sqrt{2}}, \frac{-1-i}{\sqrt{2}}\right)$ are $\varphi_{D, 1}=\frac{3 \pi}{4}$ and $\varphi_{D, 2}=\frac{5 \pi}{4}$.
The parameter vector for this matrix is therefore $\boldsymbol{\varphi}=\left[\boldsymbol{\varphi}_{L^{\prime}}^{\prime}, \boldsymbol{\varphi}_{R}^{\prime}, \boldsymbol{\varphi}_{D}^{\prime}\right]^{\prime}=\left[[0,0,0,0],\left[\frac{\pi}{4}, 0\right],\left[\frac{3 \pi}{4}, \frac{5 \pi}{4}\right]\right]^{\prime}$, with $\boldsymbol{\varphi}=C_{U}^{-1}(C)$.

Components of the Parameter Vector
Based on the results of the preceding sections we can now describe the parameter vectors for the general case. The dimensions of the parameter vectors of the respective blocks of the system matrices $(\mathcal{A}, \mathcal{B}, \mathcal{C})$ depend on the multi-index $\Gamma$, consisting of the state space unit root structure $\Omega_{S}$, the structure indices $p$ and the Kronecker indices $\alpha_{\bullet}$ for the stable subsystem. A parameterization of the set of all systems in canonical form with given multi-index $\Gamma$ for the $\operatorname{MFI}(1)$ case, therefore, combines the following components:

- $\boldsymbol{\theta}_{B, f}:=\left[\boldsymbol{\theta}_{B, f, 1}^{\prime}, \ldots, \boldsymbol{\theta}_{B, f, l}^{\prime}\right]^{\prime} \in \Theta_{B, f}=\mathbb{R}^{d_{B, f}}$, with:

$$
\boldsymbol{\theta}_{B, f, k}:=\left\{\begin{array}{lr}
{\left[b_{1, p_{1}^{k}+1^{k}}^{k}, b_{1, p_{1}^{k}+2^{\prime}}^{k} \ldots, b_{1, s^{\prime}}^{k}, b_{2, p_{2}^{k}+1^{\prime}}^{k}, \ldots, b_{d_{1}^{k}, s}^{k}\right]^{\prime}} & \text { for } \omega_{k} \in\{0, \pi\}, \\
{\left[\mathcal{R}\left(b_{1, p_{1}^{k}+1}^{k}\right), \mathcal{I}\left(b_{1, p_{1}^{k}+1}^{k}\right), \mathcal{R}\left(b_{1, p_{1}^{k}+2}^{k}\right), \ldots, \mathcal{I}\left(b_{1, s}^{k}\right), \mathcal{R}\left(b_{2, p_{2}^{k}+1}^{k}\right), \ldots, \mathcal{I}\left(b_{d_{1}^{k}, s}^{k}\right)\right]^{\prime}} \\
& \text { for } 0<\omega_{k}<\pi
\end{array}\right.
$$

for $k=1, \ldots, l$, with $p_{j}^{k}$ denoting the $j$-th entry of the structure indices $p$ corresponding to $\mathcal{B}_{k}$. The vectors $\theta_{B, f, k}$ contain the real and imaginary parts of free entries in $\mathcal{B}_{k}$ not restricted by the p.u.t. structures.

- $\boldsymbol{\theta}_{B, p}:=\left[\boldsymbol{\theta}_{B, p, 1}^{\prime}, \ldots, \boldsymbol{\theta}_{B, p, l}^{\prime}\right]^{\prime} \in \Theta_{B, p}=\mathbb{R}_{+}^{d_{B, p}}:$ The vectors $\boldsymbol{\theta}_{B, p, k}:=\left[b_{1, p_{1}^{k}}^{k} \ldots, b_{d_{1}^{k}, p_{d_{1}^{k}}^{k}}^{k}\right]^{\prime}$ contain the entries in $\mathcal{B}_{k}$ restricted by the p.u.t. structures to be positive reals.
- $\quad \boldsymbol{\theta}_{C, E}:=\left[\boldsymbol{\theta}_{C, E, 1}^{\prime}, \ldots, \boldsymbol{\theta}_{C, E, l}^{\prime}\right]^{\prime} \in \Theta_{C, E} \subset \mathbb{R}^{d_{C, E}}$ : The parameters for the matrices $\mathcal{C}_{k}$ as discussed in Lemma 1 and Lemma 2.
- $\quad \theta_{\bullet} \in \Theta_{\bullet, \alpha} \subset \mathbb{R}^{d_{\bullet}}:$ The parameters for the stable subsystem in echelon canonical form for Kronecker indices $\alpha$.

Example 7. Consider an MFI(1) process with $\Omega_{S}=\left((0,2),\left(\frac{\pi}{2}, 2\right)\right), p=[1,3,1,2,1,2]^{\prime}, n_{\bullet}=0$, and system matrices

$$
\begin{aligned}
\mathcal{A} & =\operatorname{diag}(1,1, i, i,-i,-i), \\
\mathcal{B} & =\left[\begin{array}{ccc}
1 & -1 & 2 \\
0 & 0 & 2 \\
\hline 1 & 1+i & 1-i \\
0 & 2 & i \\
\hline 1 & 1-i & 1+i \\
0 & 2 & -i
\end{array}\right], \quad \mathcal{C}=\left[\begin{array}{cc|cc|cc}
0 & \frac{1}{\sqrt{2}} & \frac{1-i}{2} & \frac{1-i}{2} & \frac{1+i}{2} & \frac{1+i}{2} \\
\frac{-1}{\sqrt{2}} & \frac{1}{2} & \frac{1+i}{2} & \frac{-1-i}{2} & \frac{1-i}{2} & \frac{-1+i}{2} \\
\frac{1}{\sqrt{2}} & \frac{1}{2} & 0 & 0 & 0 & 0
\end{array}\right],
\end{aligned}
$$

in canonical form. For this example it holds that $\boldsymbol{\theta}_{B, f}=[[-1,2],[1,1,1,-1,0,1]]^{\prime}, \boldsymbol{\theta}_{B, p}=[[1,2],[1,2]]$ and

$$
\theta_{C, E}=\left[\left[\left[\frac{\pi}{4}, \frac{7 \pi}{4}\right],\left[\frac{\pi}{2}\right]\right],\left[[0,0,0,0],\left[\frac{\pi}{4}, 0\right],\left[\frac{3 \pi}{4}, \frac{5 \pi}{4}\right]\right]\right]^{\prime}
$$

with parameter values corresponding to the $C$-blocks collected in $\boldsymbol{\theta}_{C, E}$ considered in Examples 5 and 6 .

### 3.2. The Parameterization in the $I(2)$ Case

The canonical form provided above for the general case has the following form for $\mathrm{I}(2)$ processes with unit root structure $\Omega_{s}=\left(\left(0, d_{1}^{1}, d_{2}^{1}\right)\right)$ :

$$
\mathcal{A}=\left[\begin{array}{cccc}
I_{d_{1}^{1}} & I_{d_{1}^{1}} & 0 & 0 \\
0 & I_{d_{1}^{1}} & 0 & 0 \\
0 & 0 & I_{d_{2}^{1}-d_{1}^{1}} & 0 \\
0 & 0 & 0 & \mathcal{A}_{\bullet}
\end{array}\right], \quad \mathcal{B}=\left[\begin{array}{c}
\mathcal{B}_{1,1} \\
\mathcal{B}_{1,2,1} \\
\mathcal{B}_{1,2,2} \\
\mathcal{B}_{\bullet}
\end{array}\right], \mathcal{C}=\left[\begin{array}{llll}
\mathcal{C}_{1,1}^{E} & \mathcal{C}_{1,2}^{G} & \mathcal{C}_{1,2}^{E} & \mathcal{C}_{\bullet}
\end{array}\right]
$$

where $0<d_{1}^{1} \leq d_{2}^{1} \leq s, \mathcal{B}_{1,2,1}$ and $\mathcal{B}_{1,2,2}$ are p.u.t., $\mathcal{C}_{1,1}^{E} \in O_{s, d_{1}^{1}}, \mathcal{C}_{1,2}^{E} \in O_{s, d_{2}^{1}-d_{1}^{1}},\left(\mathcal{C}_{1,1}^{E}\right)^{\prime} \mathcal{C}_{1,2}^{E}=0_{d_{1}^{1} \times d_{2}^{1}}$, $\left(C_{1,1}^{E}\right)^{\prime} C_{1,2}^{G}=0_{d_{1}^{1} \times d_{1}^{1}},\left(C_{1,2}^{E}\right)^{\prime} C_{1,2}^{G}=0_{\left(d_{2}^{1}-d_{1}^{1}\right) \times d_{1}^{1}}$ and $\left(\mathcal{A} \bullet, \mathcal{B}_{\bullet}, \mathcal{C}_{\bullet}\right)$ is in echelon canonical form with Kronecker indices $\alpha_{0}$. All matrices are real valued.

The parameterizations of the p.u.t. matrices $\mathcal{B}_{1,2,1}$ and $\mathcal{B}_{1,2,2}$ are as discussed above. The entries of $\mathcal{B}_{1,1}$ are unrestricted and thus included in the parameter vector $\boldsymbol{\theta}_{B, f}$ containing also the free entries in $\mathcal{B}_{1,2,1}$ and $\mathcal{B}_{1,2,2}$. The subsystem $\left(\mathcal{A}_{\bullet}, \mathcal{B}_{\bullet}, \mathcal{C}_{\bullet}\right)$ is parameterized using the echelon canonical form.

The parameterization of $\mathcal{C}_{1,1}^{E} \in O_{s, d_{1}^{1}}$ proceeds as in the $\operatorname{MFI}(1)$ case, using $C_{O}^{-1}\left(\mathcal{C}_{1,1}^{E}\right)$. The parameterization of $\mathcal{C}_{1,2}^{E}$ has to take the restriction of orthogonality of $\mathcal{C}_{1,2}^{E}$ to $\mathcal{C}_{1,1}^{E}$ into account, thus the set to be parameterized is given by

$$
\begin{equation*}
O_{s, d_{2}^{1}-d_{1}^{1}}\left(\mathcal{C}_{1,1}^{E}\right):=\left\{\mathcal{C}_{1,2}^{E} \in \mathbb{R}^{s \times\left(d_{2}^{1}-d_{1}^{1}\right)} \mid\left(\mathcal{C}_{1,1}^{E}\right)^{\prime} \mathcal{C}_{1,2}^{E}=0_{d_{1}^{1} \times\left(d_{2}^{1}-d_{1}^{1}\right)}\left(\mathcal{C}_{1,2}^{E}\right)^{\prime} \mathcal{C}_{1,2}^{E}=I_{d_{2}^{1}-d_{1}^{1}}\right\} \tag{13}
\end{equation*}
$$

The parameterization of this set again uses real Givens rotations. For $\mathcal{C} \in O_{s, d_{2}^{1}-d_{1}^{1}}\left(\mathcal{C}_{1,1}^{E}\right)$ it follows that $R_{L}\left(\boldsymbol{\theta}_{L}\right) \mathcal{C}=\left[0_{d_{1}^{1} \times\left(d_{2}^{1}-d_{1}^{1}\right)}^{\prime} \tilde{\mathcal{C}}^{\prime}\right]^{\prime}$ for a matrix $\tilde{\mathcal{C}}$ such that $\tilde{\mathcal{C}}^{\prime} \tilde{\mathcal{C}}=I_{d_{2}^{1}-d_{1}^{1}}$ with $R_{L}\left(\boldsymbol{\theta}_{L}\right)$ corresponding to $\mathcal{C}_{1,1}^{E}$. The matrix $\tilde{\mathcal{C}}$ is parameterized as discussed in Lemma 1.

Corollary 1 (Properties of the parameterization of $O_{s, d_{2}^{1}-d_{1}^{1}}\left(\mathcal{C}_{1,1}^{E}\right)$ ). Define for $d_{1}^{1}<d_{2}^{1} \leq s$ a mapping $\tilde{\boldsymbol{\theta}} \rightarrow C_{O, d_{2}^{1}-d_{1}^{1}}\left(\tilde{\boldsymbol{\theta}} ; C_{1,1}^{E}\right)$ from $\Theta_{O, d_{2}^{1}}^{\mathbb{R}}:=[0,2 \pi)^{\left(d_{2}^{1}-d_{1}^{1}\right)\left(s-d_{2}^{1}\right)} \times[0,2 \pi)^{\left(d_{2}^{1}-d_{1}^{1}\right)\left(d_{2}^{1}-d_{1}^{1}-1\right) / 2} \rightarrow O_{s, d_{2}^{1}-d_{1}^{1}}\left(\mathcal{C}_{1,1}^{E}\right) b y$

$$
\left.C_{O, d_{2}^{1}-d_{1}^{1}} \tilde{\boldsymbol{\theta}} ; \mathcal{C}_{1,1}^{E}\right):=R_{L}\left(\boldsymbol{\theta}_{L}\right)^{\prime}\left[\begin{array}{c}
0_{d_{1}^{1} \times\left(d_{2}^{1}-d_{1}^{1}\right)} \\
C_{O}(\tilde{\boldsymbol{\theta}})
\end{array}\right]
$$

where $\boldsymbol{\theta}_{L}$ denotes the parameter values corresponding to $\left[\boldsymbol{\theta}_{L}^{\prime}, \boldsymbol{\theta}_{R}^{\prime}\right]^{\prime}=C_{O}^{-1}\left(\mathcal{C}_{1,1}^{E}\right)$ as defined in Lemma 1. The following properties hold:
(i) $O_{s, d_{2}^{1}-d_{1}^{1}}\left(\mathcal{C}_{1,1}^{E}\right)$ is closed and bounded.
(ii) The mapping $\left.C_{O, d_{2}^{1}-d_{1}^{1}} \tilde{\boldsymbol{\theta}} ; \mathcal{C}_{1,1}^{E}\right)$ is infinitely often differentiable.

For $d_{2}^{1}<s$, it holds
(iii) For every $\mathcal{C}_{1,2}^{E} \in O_{s, d_{2}^{1}-d_{1}^{1}}\left(\mathcal{C}_{1,1}^{E}\right)$ there exists a vector $\tilde{\boldsymbol{\theta}}=\left[\tilde{\boldsymbol{\theta}}_{L}^{\prime}, \tilde{\boldsymbol{\theta}}_{R}^{\prime}\right]^{\prime} \in \Theta_{O, d_{2}^{1}-d_{1}^{1}}^{\mathbb{R}}$ such that

$$
\mathcal{C}_{1,2}^{E}=C_{O, d_{2}^{1}-d_{1}^{1}}\left(\tilde{\boldsymbol{\theta}} ; \mathcal{C}_{1,1}^{E}\right)=R_{L}\left(\boldsymbol{\theta}_{L}\right)^{\prime}\left[\begin{array}{c}
0_{d_{1}^{1} \times\left(d_{2}^{1}-d_{1}^{1}\right)} \\
\left.R_{L}\left(\tilde{\boldsymbol{\theta}}_{L}\right)^{\prime}\left[\begin{array}{c}
I_{d_{2}^{1}-d_{1}^{1}} \\
0_{\left(s-d_{2}^{1}\right) \times\left(d_{2}^{1}-d_{1}^{1}\right)}
\end{array}\right] R_{R}\left(\tilde{\boldsymbol{\theta}}_{R}\right)\right] .
\end{array}\right.
$$

The algorithm discussed above Lemma 1 defines the inverse mapping $C_{O, d_{2}^{1}-d_{1}^{1}}^{-1}$.
(iv) The inverse mapping $C_{O, d_{2}^{1}-d_{1}^{1}}^{-1}\left(\because \mathcal{C}_{1,1}^{E}\right)$ —the parameterization of $O_{s, d_{2}^{1}-d_{1}^{1}}\left(\mathcal{C}_{1,1}^{E}\right)$ —is infinitely often differentiable on the pre-image of the interior of $\Theta_{O, d_{2}^{1}-d_{1}^{1}}^{\mathbb{R}}$. This is an open and dense subset of $O_{s, d_{2}^{1}-d_{1}^{1}}\left(\mathcal{C}_{1,1}^{E}\right)$.

For $d_{2}^{1}=s$, it holds that
(v) $O_{s, s-d_{1}^{1}}\left(\mathcal{C}_{1,1}^{E}\right)$ is a disconnected space with two disjoint non-empty closed subsets:

$$
\begin{aligned}
& O_{s, s-d_{1}^{1}}^{+}\left(\mathcal{C}_{1,1}^{E}\right):= \\
& \quad\left\{\mathcal{C}_{1,2}^{E} \in \mathbb{R}^{s \times\left(s-d_{1}^{1}\right)} \mid\left(\mathcal{C}_{1,1}^{E}\right)^{\prime} \mathcal{C}_{1,2}^{E}=0_{d_{1}^{1} \times\left(s-d_{1}^{1}\right)},\left(\mathcal{C}_{1,2}^{E}\right)^{\prime} \mathcal{C}_{1,2}^{E}=I_{s-d_{1}^{1}}, \operatorname{det}\left(\left[\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{E}\right]\right)=1\right\}, \\
& O_{s, s-d_{1}^{1}}^{-}\left(\mathcal{C}_{1,1}^{E}\right):= \\
& \quad\left\{\mathcal{C}_{1,2}^{E} \in \mathbb{R}^{s \times\left(s-d_{1}^{1}\right)} \mid\left(\mathcal{C}_{1,1}^{E}\right)^{\prime} \mathcal{C}_{1,2}^{E}=0_{d_{1}^{1} \times\left(s-d_{1}^{1}\right)^{\prime}}\left(\mathcal{C}_{1,2}^{E}\right)^{\prime} \mathcal{C}_{1,2}^{E}=I_{s-d_{1}^{1}}, \operatorname{det}\left(\left[\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{E}\right]\right)=-1\right\} .
\end{aligned}
$$

(vi) For every $O_{s, s-d_{1}^{1}}^{+}\left(\mathcal{C}_{1,1}^{E}\right)$ there exists a vector $\tilde{\boldsymbol{\theta}} \in \Theta_{O, d_{2}^{1}-d_{1}^{1}}^{\mathbb{R}}$ such that

$$
\mathcal{C}_{1,2}^{E}=C_{O, s-d_{1}}\left(\tilde{\boldsymbol{\theta}} ; \mathcal{C}_{1,1}^{E}\right)=R_{R}\left(\tilde{\boldsymbol{\theta}}_{R}\right)
$$

Steps 1-4 of the algorithm discussed above Lemma 1 define the inverse mapping $C_{O, s-d_{1}^{1}}^{-1}\left(\cdot ; \mathcal{C}_{1,1}^{E}\right)$ : $O_{s, s-d_{1}^{1}}^{+}\left(\mathcal{C}_{1,1}^{E}\right) \rightarrow \Theta_{O, s-d_{1}^{1}}^{\mathbb{R}}$.
(vii) Define $v:=[\pi, \ldots, \pi]^{\prime} \in \mathbb{R}^{\left(s-d_{1}^{1}\right)\left(s-d_{1}^{1}-1\right) / 2}$. Then a parameterization of $O_{s, s-d_{1}^{1}}\left(\mathcal{C}_{1,1}^{E}\right)$ is given by

$$
C_{O, s-d_{1}^{1}}^{ \pm}\left(\mathcal{C}_{1,2}^{E} ; \mathcal{C}_{1,1}^{E}\right)= \begin{cases}v+C_{O, s-d_{1}^{1}}^{-1}\left(\mathcal{C}_{1,2}^{E} ; \mathcal{C}_{1,1}^{E}\right) & \text { if } C \in O_{s, s-d_{1}^{1}}^{+}\left(\mathcal{C}_{1,1}^{E}\right) \\ -\left(v+C_{O, s-d_{1}^{1}}^{-1}\left(\mathcal{C}_{1,2}^{E} I_{s-d_{1}^{1}}^{-} ; \mathcal{C}_{1,1}^{E}\right)\right) & \text { if } C \in O_{s, s-d_{1}^{1}}^{-}\left(\mathcal{C}_{1,1}^{E}\right)\end{cases}
$$

The parameterization is infinitely often differentiable with infinitely often differentiable inverse on an open and dense subset of $O_{s, s}$.

The proof of Corollary 1 uses the same arguments as the proof of Lemma 1 and is, therefore, omitted. It remains to provide a parameterization for $\mathcal{C}_{1,2}^{G}$ restricted to be orthogonal to both $\mathcal{C}_{1,1}^{E}$ and $\mathcal{C}_{1,2}^{E}$. Thus, the set to be parametrized is given by

$$
O_{s, G}\left(\mathcal{C}_{1,1}^{E} \mathcal{C}_{1,2}^{E}\right):=\left\{\mathcal{C}_{1,2}^{G} \in \mathbb{R}^{s \times d_{1}^{1}} \mid\left(\mathcal{C}_{1,1}^{E}\right)^{\prime} \mathcal{C}_{1,2}^{G}=0_{d_{1}^{1} \times d_{1}^{1}}\left(\mathcal{C}_{1,2}^{E}\right)^{\prime} \mathcal{C}_{1,2}^{G}=0_{\left(d_{2}^{1}-d_{1}^{1}\right) \times d_{1}^{1}}\right\} .
$$

The parameterization of $O_{s, G}\left(\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{E}\right)$ is straightforward: Left multiplication of $\mathcal{C}_{1,2}^{G}$ with $R_{L}\left(\boldsymbol{\theta}_{L}\right)$ as defined in Lemma 1 and of the lower $\left(s-d_{1}^{1}\right) \times d_{1}^{1}$ - block with $R_{L}\left(\tilde{\boldsymbol{\theta}}_{L}\right)$ as defined in Corollary 1 transforms the upper $d_{2}^{1} \times d_{1}^{1}$-block to zero and collects the free parameters in the lower $\left(s-d_{2}^{1}\right) \times$ $d_{1}^{1}$-block. Clearly, this is a bijective and infinitely often differentiable mapping on $O_{s, G}\left(\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{E}\right)$ and thus a useful parameterization, since the matrix $\mathcal{C}_{1,2}^{G}$ is only multiplied with two constant invertible matrices. The entries of the matrix product are then collected in a parameter vector as shown in Corollary 2.

Corollary 2 (Properties of the parameterization of $O_{s, G}\left(\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{E}\right)$ ). Define for given matrices $\mathcal{C}_{1,1}^{E} \in O_{s, d_{1}^{1}}$ and $\mathcal{C}_{1,2}^{E} \in O_{s, 1_{2}^{1}-d_{1}^{1}}\left(\mathcal{C}_{1,1}^{E}\right)$ a mapping $\lambda \rightarrow C_{O, G}\left(\lambda ; \mathcal{C}_{1,1}^{E} \mathcal{C}_{1,2}^{E}\right)$ from $\mathbb{R}^{d_{1}^{1}\left(s-d_{2}^{1}\right)} \rightarrow O_{s, G}\left(\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{E}\right)$ by

$$
\mathcal{C}_{O, G}\left(\lambda ; \mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{E}\right):=R_{L}\left(\boldsymbol{\theta}_{L}\right)^{\prime}\left[R_{L}\left(\tilde{\boldsymbol{\theta}}_{L}\right)^{\prime}\left[\begin{array}{ccc}
0_{d_{1}^{1} \times d_{1}^{1}} & \\
0_{\left(d_{2}^{1}-d_{1}^{1}\right) \times 1} & \cdots & 0_{\left(d_{2}^{1}-d_{1}^{1}\right) \times 1} \\
\lambda_{1} & \cdots & \lambda_{d_{1}^{1}} \\
\lambda_{d_{1}^{1}+1} & \cdots & \lambda_{2 d_{1}^{1}} \\
\vdots & & \vdots \\
\lambda_{d_{1}^{1}\left(s-d_{2}^{1}-1\right)+1} & \cdots & \lambda_{d_{1}^{1}\left(s-d_{2}^{1}\right)}
\end{array}\right]\right]
$$

where $\boldsymbol{\theta}_{L}$ denotes the parameter values corresponding to $\left[\boldsymbol{\theta}_{L}^{\prime}, \boldsymbol{\theta}_{R}^{\prime}\right]^{\prime}=C_{O}^{-1}\left(\mathcal{C}_{1,1}^{E}\right)$ as defined in Lemma 1 and $\tilde{\boldsymbol{\theta}}_{L}$ denotes the parameter values corresponding to $\left[\tilde{\boldsymbol{\theta}}_{L}^{\prime}, \tilde{\boldsymbol{\theta}}_{R}^{\prime}\right]^{\prime}=C_{O, d_{2}^{1}-d_{1}^{1}}^{-1}\left(\mathcal{C}_{1,2}^{E} ; \mathcal{C}_{1,1}^{E}\right)$ as defined in Corollary 1. The set $O_{s, G}\left(\mathcal{C}_{1,1}^{E} \mathcal{C}_{1,2}^{E}\right)$ is closed and both $C_{O, G}$ as well as $C_{O, G}^{-1}(\cdot)$-the parameterization of $O_{s, G}\left(\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{E}\right)$-are infinitely often differentiable.

Components of the Parameter Vector
In the $I(2)$ case, the multi-index $\Gamma$ contains the state space unit root structure $\Omega_{S}=\left(\left(0, d_{1}^{1}, d_{2}^{1}\right)\right)$, the structure indices $p \in \mathbb{N}_{0}^{d_{1}^{1}+d_{2}^{1}}$, encoding the p.u.t. structures of $\mathcal{B}_{1,2,1}$ and $\mathcal{B}_{1,2,2}$, and the Kronecker indices $\alpha_{\bullet}$ for the stable subsystem. The parameterization of the set of all systems in canonical form with given multi-index $\Gamma$ for the $I(2)$ case uses the following components:

- $\boldsymbol{\theta}_{B, f}:=\boldsymbol{\theta}_{B, f, 1} \in \Theta_{B, f}=\mathbb{R}^{d_{B, f}}$ : The vector $\boldsymbol{\theta}_{B, f, 1}$ contains the free entries in $\mathcal{B}_{1}$ not restricted by the p.u.t. structure, collected in the same order as for the matrices $\mathcal{B}_{k}$ in the $\operatorname{MFI}(1)$ case.
- $\boldsymbol{\theta}_{B, p}:=\boldsymbol{\theta}_{B, p, 1} \in \Theta_{B, p}=\mathbb{R}_{+}^{d_{B, p}}:$ The vector $\boldsymbol{\theta}_{B, p, 1}:=\left[b_{d_{1}^{1}+1, p_{d_{1}^{1}+1}^{1}}^{1}, \ldots, b_{d_{1}^{1}+d_{2}^{1}, p_{d_{1}^{1}+d_{2}^{1}}^{1}}^{1}\right]^{\prime}$ contains the entries in $\mathcal{B}_{1}$ restricted by the p.u.t. structures to be positive reals.
- $\quad \boldsymbol{\theta}_{C, E}:=\left[\boldsymbol{\theta}_{C, E, 1,1}^{\prime}, \boldsymbol{\theta}_{C, E, 1,2}^{\prime}\right]^{\prime} \in \Theta_{C, E} \subset \mathbb{R}^{d_{C, E}}$ : The parameters for the matrices $\mathcal{C}_{1,1}^{E}$ as in the $\operatorname{MFI}(1)$ case and $\mathcal{C}_{1,2}^{E}$ as discussed in Corollary 1.
- $\quad \boldsymbol{\theta}_{C, G} \in \Theta_{C, G}=\mathbb{R}^{d_{C, G}}$ : The parameters for the matrix $\mathcal{C}_{1,2}^{G}$ as discussed in Corollary 2.
- $\quad \theta_{\bullet} \in \Theta_{\bullet, \alpha} \subset \mathbb{R}^{d_{\bullet}}:$ The parameters for the stable subsystem in echelon canonical form for Kronecker indices $\alpha_{0}$.

Example 8. Consider an $I(2)$ process with $\Omega_{S}=((0,1,2)), p=[0,1,1]^{\prime}, n_{\bullet}=0$ and system matrices

$$
\mathcal{A}=\left[\begin{array}{lll}
1 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right], \quad \mathcal{B}=\left[\begin{array}{ccc}
-1 & 2 & -2 \\
1 & -1 & 3 \\
2 & 0 & 1
\end{array}\right], \quad \mathcal{C}=\left[\begin{array}{c|c|c}
0 & -1 & \frac{1}{\sqrt{2}} \\
\frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{2} \\
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{2}
\end{array}\right]
$$

In this case, $\boldsymbol{\theta}_{B, f, 1}=[-1,2,-2,-1,3,0,1]^{\prime}, \boldsymbol{\theta}_{B, p, 1}=[1,2]^{\prime}$. It follows from

$$
\begin{aligned}
& R_{3,1,2}\left(\frac{7 \pi}{4}\right) R_{3,1,3}\left(\frac{\pi}{2}\right) \mathcal{C}_{1,1}^{E}=\quad\left[\begin{array}{lll}
1 & 0 & 0
\end{array}\right]^{\prime}, \\
& R_{3,1,2}\left(\frac{7 \pi}{4}\right) R_{3,1,3}\left(\frac{\pi}{2}\right) \mathcal{C}_{1,2}^{E}=\left[\begin{array}{lll}
0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}}
\end{array}\right]^{\prime} \quad \text { and } \quad R_{2,1,2}\left(\frac{7 \pi}{4}\right)\left[\begin{array}{c}
\frac{1}{\sqrt{2}} \\
\frac{-1}{\sqrt{2}}
\end{array}\right]=\left[\begin{array}{l}
1 \\
0
\end{array}\right], \\
& R_{3,1,2}\left(\frac{7 \pi}{4}\right) R_{3,1,3}\left(\frac{\pi}{2}\right) \mathcal{C}_{1,2}^{G}=\left[\begin{array}{lll}
0 & 1 & 1
\end{array}\right]^{\prime} \quad \text { and } \quad R_{2,1,2}\left(\frac{7 \pi}{4}\right)\left[\begin{array}{l}
1 \\
1
\end{array}\right]=\left[\begin{array}{c}
0 \\
\sqrt{2}
\end{array}\right] \text {, }
\end{aligned}
$$

that $\boldsymbol{\theta}_{C, E}=\left[\boldsymbol{\theta}_{C, E, 1,1}^{\prime}, \boldsymbol{\theta}_{C, E, 1,2}\right]^{\prime}=\left[\left[\frac{\pi}{2}, \frac{7 \pi}{4}\right],\left[\frac{7 \pi}{4}\right]\right]^{\prime}$ and $\boldsymbol{\theta}_{C, G}=[\sqrt{2}]$.

### 3.3. The Parameterization in the General Case

Inspecting the canonical form shows that all relevant building blocks are already present in the $\operatorname{MFI}(1)$ and the $\mathrm{I}(2)$ cases and can be combined to deal with the general case: The entries in $\mathcal{B}_{u}$ are either unrestricted or follow restrictions according to given structure indices $p$, and the parameter space is chosen accordingly, as discussed for the $\operatorname{MFI}(1)$ and $\mathrm{I}(2)$ cases. The restrictions on the matrices $\mathcal{C}_{u}$ and its blocks $\mathcal{C}_{k}$ require more sophisticated parameterizations of parts of unitary or orthonormal matrices as well as of orthogonal complements. These are dealt with in Lemmas 1 and 2 and Corollaries 1 and 2 above. The extension of Corollaries 1 and 2 to complex matrices and to matrices which are orthogonal to a larger number of blocks of $\mathcal{C}_{k}$ is straightforward.

The following theorem characterizes the properties of parameterizations for sets $M_{\Gamma}$ of transfer functions with (general) multi-index $\Gamma$ and describes the relations between sets of transfer functions and the corresponding sets $\Delta_{\Gamma}$ of triples $(\mathcal{A}, \mathcal{B}, \mathcal{C})$ of system matrices in canonical form, defined below. Discussing the continuity and differentiability of mappings on sets of transfer functions and on sets of matrix triples also requires the definition of a topology on both sets.

## Definition 8.

(i) The set of transfer functions of order $n, M_{n}$, is endowed with the pointwise topology $T_{p t}$ : First, identify transfer functions with their impulse response sequences. Then, a sequence of transfer functions $k_{i}(z)=$ $I_{s}+\sum_{j=1}^{\infty} K_{j, i} z^{j}$ converges in $T_{p t}$ to $k_{0}(z)=I_{s}+\sum_{j=1}^{\infty} K_{j, 0} z^{j}$ if and only if for every $j \in \mathbb{N}$ it holds that $K_{j, i} \xrightarrow{i \rightarrow \infty} K_{j, 0}$.
(ii) The set of all triples $(\mathcal{A}, \mathcal{B}, \mathcal{C})$ in canonical form corresponding to transfer functions with multi-index $\Gamma$ is called $\Delta_{\Gamma}$. The set $\Delta_{\Gamma}$ is endowed with the topology corresponding to the distance $d\left(\left(A_{1}, B_{1}, C_{1}\right),\left(A_{2}, B_{2}, C_{2}\right)\right):=\left\|A_{1}-A_{2}\right\|_{F r}+\left\|B_{1}-B_{2}\right\|_{F r}+\left\|C_{1}-C_{2}\right\|_{F r}$.

Please note that in the definition of the pointwise topology convergence does not need to be uniform in $j$ and moreover, the power series coefficients do not need to converge to zero for $j \rightarrow \infty$ and hence the concept can also be used for unstable systems.

Theorem 2. The set $M_{n}$ can be partitioned into pieces $M_{\Gamma}$, where $\Gamma:=\left\{\Omega_{S}, p, \alpha_{\bullet}\right\}$, i.e.,

$$
M_{n}=\bigcup_{\Gamma=\left\{\Omega_{S}, p, \alpha_{\bullet}\right\} \mid n_{u}\left(\Omega_{S}\right)+n_{\bullet}\left(\alpha_{\bullet}\right)=n} M_{\Gamma},
$$

where $n_{u}\left(\Omega_{S}\right):=\sum_{k=1}^{l} \sum_{j=1}^{h_{k}} d_{j}^{k} \delta_{k}$, with $\delta_{k}=1$ for $\omega_{k} \in\{0, \pi\}$ and $\delta_{k}=2$ for $0<\omega_{k}<\pi$ is the state dimension of the unstable subsystem $\left(\mathcal{A}_{u}, \mathcal{B}_{u}, \mathcal{C}_{u}\right)$ with state space unit root structure $\Omega_{S}$ and $n_{\bullet}\left(\alpha_{\bullet}\right):=$ $\sum_{i=1}^{s} \alpha_{\bullet, i}$ is the state dimension of the stable subsystem with Kronecker indices $\alpha_{\bullet}=\left(\alpha_{\bullet}, 1, \ldots, \alpha_{\bullet}, s\right), \alpha_{\bullet}, i \in \mathbb{N}_{0}$. For every multi-index $\Gamma$ there exists a parameter space $\Theta_{\Gamma} \subset \mathbb{R}^{d(\Gamma)}$ for some integer $d(\Gamma)$, endowed with the Euclidean norm, and a function $\phi_{\Gamma}: \Delta_{\Gamma} \rightarrow \Theta_{\Gamma}$, such that for every $(\mathcal{A}, \mathcal{B}, \mathcal{C}) \in \Delta_{\Gamma}$ the parameter vector $\theta:=\phi_{\Gamma}(\mathcal{A}, \mathcal{B}, \mathcal{C}) \in \Theta_{\Gamma}$ is composed of:

- The parameter vector $\boldsymbol{\theta}_{B, f}=\left[\boldsymbol{\theta}_{B, f, 1}^{\prime}, \ldots, \boldsymbol{\theta}_{B, f, l}^{\prime}\right]^{\prime} \in \Theta_{B, f}=\mathbb{R}^{d_{B, f}}$, collecting the (real and imaginary parts of) non-restricted entries in $\mathcal{B}_{k}, k=1, \ldots$, l as described in the $M F I(1)$ case.
- The parameter vector $\boldsymbol{\theta}_{B, p}=\left[\boldsymbol{\theta}_{B, p, 1}^{\prime}, \ldots, \boldsymbol{\theta}_{B, p, l}^{\prime}\right]^{\prime} \in \Theta_{B, p}=\mathbb{R}_{+}^{d_{B, p}}$, collecting the entries in $\mathcal{B}_{k}, k=1, \ldots, l$, restricted by the p.u.t. forms to be positive reals in a similar fashion as described for $\mathcal{B}_{1}$ in the $I(2)$ case.
- The parameter vector $\boldsymbol{\theta}_{C, E}=\left[\boldsymbol{\theta}_{C, E, 1}^{\prime}, \ldots, \boldsymbol{\theta}_{C, E, l}^{\prime}\right]^{\prime} \in \Theta_{C, E} \subset \mathbb{R}^{d_{C, E}}, \boldsymbol{\theta}_{C, E, k}=\left[\boldsymbol{\theta}_{C, E, k, 1}^{\prime}, \ldots, \boldsymbol{\theta}_{C, E, k, h_{k}}^{\prime}\right]^{\prime}$ collecting the parameters $\boldsymbol{\theta}_{C, E, k, j}$ for all blocks $\mathcal{C}_{k, j^{\prime}}^{E} k=1, \ldots, l$ and $j=1, \ldots, h_{k}$, obtained using Givens rotations (see Lemmas 1 and 2 and Corollary 1 and its extension to complex matrices).
- The parameter vector $\boldsymbol{\theta}_{C, G}=\left[\boldsymbol{\theta}_{C, G, 1}^{\prime}, \ldots, \boldsymbol{\theta}_{C, G, l}^{\prime}\right]^{\prime} \in \Theta_{C, G}=\mathbb{R}^{d, G}, \boldsymbol{\theta}_{C, G, k}=\left[\boldsymbol{\theta}_{C, G, k, 2}^{\prime}, \ldots, \boldsymbol{\theta}_{C, G, k, h_{k}}^{\prime}\right]^{\prime}$ collecting the parameters $\boldsymbol{\theta}_{C, G, k, j}$ (real and imaginary parts for complex roots) for $\mathcal{C}_{k, j}^{G}, k=1, \ldots, l$ and $j=$ $2, \ldots, h_{k}$, subject to the orthogonality restrictions (see Corollary 2 and its extension to complex matrices).
- The parameter vector $\theta_{\bullet} \in \Theta \bullet \subset \mathbb{R}^{d \bullet}$ collecting the free entries in echelon canonical form with Kronecker indices $\alpha_{0}$.
(i) The mapping $\psi_{\Gamma}: M_{\Gamma} \rightarrow \Delta_{\Gamma}$ that attaches a triple $(\mathcal{A}, \mathcal{B}, \mathcal{C})$ in canonical form to a transfer function in $M_{\Gamma}$ is continuous. It is the inverse (restricted to $M_{\Gamma}$ ) of the $T_{p t}$-continuous function $\pi:(A, B, C) \mapsto$ $k(z)=I_{s}+z C\left(I_{n}-z A\right)^{-1} B$.
(ii) Every parameter vector $\boldsymbol{\theta}=\left[\boldsymbol{\theta}_{B, f}^{\prime}, \boldsymbol{\theta}_{B, p}^{\prime}, \boldsymbol{\theta}_{C, E}^{\prime}, \boldsymbol{\theta}_{C, G}^{\prime}, \boldsymbol{\theta}_{\bullet}^{\prime}\right]^{\prime} \in \Theta_{\Gamma} \subset \Theta_{B, f} \times \Theta_{B, p} \times \Theta_{C, E} \times$ $\Theta_{C, G} \times \Theta_{\text {. }}$ corresponds to a triple $(\mathcal{A}(\boldsymbol{\theta}), \mathcal{B}(\boldsymbol{\theta}), \mathcal{C}(\boldsymbol{\theta})) \in \Delta_{\Gamma}$ and a transfer function $k(z)=$ $\pi(\mathcal{A}(\boldsymbol{\theta}), \mathcal{B}(\boldsymbol{\theta}), \mathcal{C}(\boldsymbol{\theta})) \in M_{\Gamma}$. The mapping $\phi_{\Gamma}^{-1}: \boldsymbol{\theta} \rightarrow(\mathcal{A}(\boldsymbol{\theta}), \mathcal{B}(\boldsymbol{\theta}), \mathcal{C}(\boldsymbol{\theta}))$ is continuous on $\Theta_{\Gamma}$.
(iii) For every multi-index $\Gamma$ the set of points in $\Delta_{\Gamma}$, where the mapping $\phi_{\Gamma}$ is continuous, is open and dense in $\Delta_{\Gamma}$.

As mentioned in Section 2, the parameterization of $\Phi$ is straightforward. The $s \times m$ entries of $\Phi$ are collected in a parameter vector $d$. Thus, there is a one-to-one correspondence between state space realizations $(\mathcal{A}, \mathcal{B}, \mathcal{C}, \Phi) \in \Delta_{\Gamma} \times \mathbb{R}^{s \times m}$ and parameter vectors $\boldsymbol{\tau}=\left[\boldsymbol{\theta}^{\prime}, \boldsymbol{d}^{\prime}\right]^{\prime} \in \Theta_{\Gamma} \times \mathbb{R}^{s m}$. The same holds true for parameters used for the symmetric, positive definite innovation matrix $\Sigma \in \mathbb{R}^{s \times s}$ obtained, e.g., from a lower triangular Cholesky factor of $\Sigma$.

## 4. The Topological Structure

The parameterization of $M_{n}$ in Theorem 2 partitions $M_{n}$ into subsets $M_{\Gamma}$ for a selection of multi-indices $\Gamma$. To every multi-index $\Gamma$ there exists a corresponding associated parameter set $\Theta_{\Gamma}$. Thus, in practical applications, maximizing the pseudo likelihood requires choosing the multi-index $\Gamma$. Maximizing the pseudo likelihood over the set $M_{\Gamma}$ effectively amounts to including also all elements in the closure of $M_{\Gamma}$, because of continuity of the parameterization. It is thus necessary to characterize the closures of the sets $M_{\Gamma}$.

Moreover, maximizing the pseudo likelihood function over all possible multi-indices is time-consuming and not desirable. Fortunately, the results discussed below show that there exists a generic multi-index $\Gamma_{g}$ such that $M_{n} \subset \overline{M_{\Gamma_{g}}}$. This generic choice corresponds to the set of all stable systems of order $n$ corresponding to the generic neighborhood of the echelon canonical form. This multi-index, therefore, is a natural starting point for estimation.

However, in particular for hypotheses testing, it will be necessary to maximize the pseudo likelihood over sets of transfer functions of order $n$ with specific state space unit root structure $\Omega_{S}$, denoted as $M\left(\Omega_{S}, n_{\bullet}\right)$ below, where $n_{\bullet}$ denotes the dimension of the stable part of the state. We show below that also in this case there exists a generic multi-index $\Gamma_{g}\left(\Omega_{S}, n_{\bullet}\right)$ such that $M\left(\Omega_{S}, n_{\bullet}\right) \subset$ $\overline{M_{\Gamma_{g}\left(\Omega_{S}, n_{\mathbf{\bullet}}\right)}}$.

The main tool to obtain these results is investigating the properties of the mappings $\psi_{\Gamma}$, that map transfer functions in $M_{\Gamma}$ to triples $(A, B, C) \in \Delta_{\Gamma}$, as well as analyzing the closures of the sets $\Delta_{\Gamma}$. The relation between parameter vectors $\theta \in \Theta_{\Gamma}$ and triples of system matrices $(A, B, C) \in$ $\Delta_{\Gamma}$ is easier to understand than the relation between $\Delta_{\Gamma}$ and $M_{\Gamma}$, due to the results of Theorem 2. Consequently, this section focuses on the relations between $\Delta_{\Gamma}$ and $M_{\Gamma}$-and their closures-for different multi-indices $\Gamma$.

To define the closures we embed the sets $\Delta_{\Gamma}$ of matrices in canonical form with multi-indices $\Gamma$ corresponding to transfer functions of order $n$ into the space $\Delta_{n}$ of all conformable complex matrix triples $(A, B, C)$ with $A \in \mathbb{C}^{n \times n}$, where additionally $\lambda_{|\max |}(A) \leq 1$. Since the elements of $\Delta_{n}$ are matrix triples, this set is isomorphic to a subset of the finite dimensional space $\mathbb{C}^{n^{2}+2 n s}$, equipped with the Euclidean topology. Please note that $\Delta_{n}$ also contains non-minimal state space realizations, corresponding to transfer functions of lower order.

Remark 16. In principle the set $\Delta_{n}$ also contains state space realizations of transfer functions $k(z)=I_{s}+$ $\sum_{j=1}^{\infty} K_{j} z^{j}$ with complex valued coefficients $K_{j}$. Since the subset of $\Delta_{n}$ of state space systems realizing transfer functions with real valued $K_{j}$ is closed in $\Delta_{n}$, realizations corresponding to transfer functions with coefficients with non-zero imaginary part are irrelevant for the analysis of the closures of the sets $\Delta_{\Gamma}$.

After investigating the closure of $\Delta_{\Gamma}$ in $\Delta_{n}$, denoted by $\overline{\Delta_{\Gamma}}$, we consider the set of corresponding transfer functions $\pi\left(\overline{\Delta_{\Gamma}}\right)$. Since we effectively maximize the pseudo likelihood over $\overline{\Delta_{\Gamma}}$, we have to understand for which multi-indices $\tilde{\Gamma}$ the set $\pi\left(\Delta_{\tilde{\Gamma}}\right)$ is a subset of $\pi\left(\overline{\Delta_{\Gamma}}\right)$. Moreover, we find a covering of $\pi\left(\overline{\Delta_{\Gamma}}\right) \subset \bigcup_{i \in \mathcal{I}} M_{\Gamma_{i}}$. This restricts the set of multi-indices $\Gamma$ that may occur as possible multi-indices of the limit of a sequence in $\pi\left(\Delta_{\Gamma}\right)$ and thus the set of transfer functions that can be obtained by maximization of the pseudo likelihood.

The sets $M_{\Gamma}$, are embedded into the vector space $M$ of all causal transfer functions $k(z)=$ $I_{s}+\sum_{j=1}^{\infty} K_{j} z^{j}$. The vector space $M$ is isomorphic to the infinite dimensional space $\Pi_{j \in \mathbb{N}} \mathbb{R}_{j}^{s \times s}$ equipped with the pointwise topology. Since, as mentioned above, maximization of the pseudo likelihood function over $M_{\Gamma}$ effectively includes $\overline{M_{\Gamma}}$, it is important to determine for any given multi-index $\Gamma$, the multi-indices $\tilde{\Gamma}$ for which the set $M_{\tilde{\Gamma}}$ is a subset of $\overline{M_{\Gamma}}$. Please note that $\overline{M_{\Gamma}}$ is not necessarily equal to $\pi\left(\overline{\Delta_{\Gamma}}\right)$. The continuity of $\pi$, as shown in Theorem 2 (i), implies the following inclusions:

$$
M_{\Gamma}=\pi\left(\Delta_{\Gamma}\right) \subset \pi\left(\overline{\Delta_{\Gamma}}\right) \subset \overline{M_{\Gamma}}
$$

In general all these inclusions are strict. For a discussion in case of stable transfer functions see Hannan and Deistler (1988, Theorem 2.5.3).

We first define a partial ordering on the set of multi-indices $\Gamma$. Subsequently we examine the closure $\bar{\Delta}_{\Gamma}$ in $\Delta_{n}$ and finally we examine the closures $\bar{M}_{\Gamma}$ in $M$.

## Definition 9.

(i) For two state space unit root structures $\Omega_{S}$ and $\tilde{\Omega}_{S}$ with corresponding matrices $\mathcal{A}_{u} \in \mathbb{C}^{n_{u} \times n_{u}}$ and $\tilde{\mathcal{A}}_{u} \in \mathbb{C}^{\tilde{n}_{u} \times \tilde{n}_{u}}$ in canonical form, it holds that $\tilde{\Omega}_{S} \leq \Omega_{S}$ if and only if there exists a permutation matrix $S$ such that

$$
S \mathcal{A}_{u} S^{\prime}=\left[\begin{array}{cc}
\tilde{\mathcal{A}}_{u} & \tilde{J}_{12} \\
0 & \tilde{J}_{2}
\end{array}\right]
$$

Moreover, $\tilde{\Omega}_{S}<\Omega_{S}$ holds if additionally $\tilde{\Omega}_{S} \neq \Omega_{S}$.
(ii) For two state space unit root structures $\Omega_{S}$ and $\tilde{\Omega}_{S}$ and dimensions of the stable subsystems $n_{\bullet}, \tilde{n} \bullet \in \mathbb{N}_{0}$ we define

$$
\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right) \leq\left(\Omega_{S}, n_{\bullet}\right) \quad \text { if and only if } \tilde{\Omega}_{S} \leq \Omega_{S}, \tilde{n}_{\bullet} \leq n_{\bullet}
$$

Strict inequality holds, if at least one of the two inequalities above holds strictly.
(iii) For two pairs $\left(\Omega_{S}, p\right)$ and $\left(\tilde{\Omega}_{S}, \tilde{p}\right)$ with corresponding matrices $\mathcal{A}_{u} \in \mathbb{C}^{n_{u} \times n_{u}}$ and $\tilde{\mathcal{A}}_{u} \in \mathbb{C}^{\tilde{n}_{u} \times \tilde{n}_{u}}$ in canonical form, it holds that $\left(\tilde{\Omega}_{S}, \tilde{p}\right) \leq\left(\Omega_{S}, p\right)$ if and only if there exists a permutation matrix $S$ such that

$$
S \mathcal{A}_{u} S^{\prime}=\left[\begin{array}{cc}
\tilde{\mathcal{A}}_{u} & \tilde{J}_{12} \\
0 & \tilde{J}_{2}
\end{array}\right], \quad S p=\left[\begin{array}{l}
p_{1} \\
p_{2}
\end{array}\right]
$$

where $p_{1} \in \mathbb{N}_{0}^{\tilde{n}_{u}}$ and $\tilde{p}$ restricts at least as many entries as $p_{1}$, i.e., $\tilde{p}_{i} \geq\left(p_{1}\right)_{i}$ holds for all $i=1, \ldots, \tilde{n}_{u}$. Moreover, $\left(\tilde{\Omega}_{S}, \tilde{p}\right)<\left(\Omega_{S}, p\right)$ holds if additionally $\left(\tilde{\Omega}_{S}, \tilde{p}\right) \neq\left(\Omega_{S}, p\right)$.
(iv) Let $\alpha_{\bullet}=\left(\alpha_{\bullet}, 1, \ldots, \alpha_{\bullet, s}\right), \alpha_{\bullet, i} \in \mathbb{N}_{0}$ and $\tilde{\alpha}_{\bullet}=\left(\tilde{\alpha}_{\bullet, 1}, \ldots, \tilde{\alpha}_{\bullet, s}\right), \tilde{\alpha}_{\bullet, i} \in \mathbb{N}_{0}$. Then $\tilde{\alpha}_{\bullet} \leq \alpha_{\bullet}$ if and only if $\tilde{\alpha}_{\bullet, i} \leq \alpha_{\bullet, i}, i=1, \ldots, s$. Moreover, $\tilde{\alpha}_{\bullet}<\alpha_{\bullet}$ holds, if at least one inequality is strict (compare Hannan and Deistler 1988, sct. 2.5).

Finally, define

$$
\tilde{\Gamma}=\left(\tilde{\Omega}_{S}, \tilde{p}, \tilde{\alpha}_{\bullet}\right) \leq \Gamma=\left(\Omega_{S}, p, \alpha_{\bullet}\right) \quad \text { if and only if }\left(\tilde{\Omega}_{S}, \tilde{p}\right) \leq\left(\Omega_{S}, p\right) \text { and } \tilde{\alpha}_{\bullet} \leq \alpha_{\bullet}
$$

Strict inequality holds, if at least one of the inequalities above holds strictly.
Please note that (i) implies that $\tilde{\Omega}_{S}$ only contains unit roots that are also contained in $\Omega_{S}$, with the integration orders $\tilde{h}_{k}$ of the unit roots in $\tilde{\Omega}_{S}$ smaller or equal to the integration orders of the respective unit roots in $\Omega_{S}$. Thus, denoting the unit root structures corresponding to $\tilde{\Omega}_{S}$ and $\Omega_{S}$ by $\tilde{\Omega}$ and $\Omega$, it follows that $\tilde{\Omega}_{S} \leq \Omega_{S}$ implies $\tilde{\Omega} \preceq \Omega$. The reverse does not hold as, e.g., for $\Omega_{S}=((0,1,1))$ (where
hence $\Omega=((0,2))$ ) and $\tilde{\Omega}_{S}=((0,2))$ (with $\tilde{\Omega}=((0,1))$ ) it holds that $\tilde{\Omega} \prec \Omega$, but neither $\tilde{\Omega}_{S} \leq \Omega_{S}$ nor $\Omega_{S} \leq \tilde{\Omega}_{S}$ holds as here

$$
\mathcal{A}_{u}=\left(\begin{array}{cc}
1 & 1 \\
0 & 1
\end{array}\right), \quad \tilde{\mathcal{A}}_{u}=\left(\begin{array}{ll}
1 & 0 \\
0 & 1
\end{array}\right)
$$

This partial ordering is convenient for the characterization of the closure of $\Delta_{\Gamma}$.

### 4.1. The Closure of $\Delta_{\Gamma}$ in $\Delta_{n}$

Please note that the block-structure of $\mathcal{A}$ implies that every system in $\Delta_{\Gamma}$ can be separated in two subsystems $\left(\mathcal{A}_{u}, \mathcal{B}_{u}, \mathcal{C}_{u}\right)$ and $\left(\mathcal{A}_{\bullet}, \mathcal{B}_{\bullet}, \mathcal{C}_{\bullet}\right)$. Define $\Delta_{\Omega_{S}, p}:=\Delta_{\left(\Omega_{S, p},\{ \}\right)}$ as the set of all state space realizations in canonical form corresponding to state space unit root structure $\Omega_{S}$, structure indices $p$ and $n_{\bullet}=0$. Analogously define $\Delta_{\alpha_{\bullet}}:=\Delta_{\left(\{ \},\{ \}, \alpha_{\bullet}\right)}$ as the set of all state space realizations in canonical form with $\Omega_{S}=\{ \}$ and Kronecker indices $\alpha_{\bullet}$. Examining $\overline{\Delta_{\Omega_{S}, p}}$ and $\overline{\Delta_{\alpha_{\bullet}}}$ separately simplifies the analysis.

### 4.1.1. The Closure of $\Delta_{\Omega_{\varsigma}, p}$

The canonical form imposes a lot of structure, i.e., restrictions on the matrices $\mathcal{A}, \mathcal{B}$ and $\mathcal{C}$. By definition $\Delta_{\Omega_{S, p}}=\Delta_{\Omega_{S}, p}^{\mathcal{A}} \times \Delta_{\Omega_{S, p}}^{\mathcal{B}} \times \Delta_{\Omega_{S}, p}^{\mathcal{C}}$ and the closures of the three matrices can be analyzed separately. $\Delta_{\Omega_{S, p}}^{\mathcal{A}}$ and $\Delta_{\Omega_{S, p}}^{\mathcal{C}}$ are very easy to investigate. The structure of $\mathcal{A}$ is fully determined by $\Omega_{S}$ and consequently $\Delta_{\Omega_{S, p}}^{\mathcal{A}}$ consists of a single matrix $\mathcal{A}$ which immediately implies that $\overline{\Delta_{\Omega_{S, p}}^{\mathcal{A}}}=\Delta_{\Omega_{S, p}}^{\mathcal{A}}$. The matrix $\mathcal{C}$, compare Theorem 1 is composed of blocks $\mathcal{C}_{k}^{E}$ that are sub-blocks of unitary (or orthonormal) matrices and blocks $\mathcal{C}_{k}^{G}$ that have to fulfill (recursive) orthogonality constraints. The corresponding sets were shown to be closed in Lemmas 1 and 2 and Corollaries 1 and 2. Thus, $\overline{\Delta_{\Omega_{S, p}}^{\mathcal{C}}}=\Delta_{\Omega_{S, p}}^{\mathcal{C}}$.

It remains to discuss $\overline{\Delta_{\Omega_{S, p}}^{\mathcal{B}}}$. The structure indices $p$ defining the p.u.t. structures of the matrices $\mathcal{B}_{k}$ restrict some entries to be positive. Combining all the parameters-unrestricted with complex values parameterized by real and imaginary part and the positive entries-into a parameter vector leads to an open sub-set of $\mathbb{R}^{m}$ for some $m$. For convergent sequences of systems with fixed $\Omega_{S}$ and $p$, limits of entries restricted to be positive may be zero. When this happens, two cases have to be distinguished. First, all p.u.t. sub-matrices still have full row rank. In this case the limiting system, $\left(\mathcal{A}_{0}, \mathcal{B}_{0}, \mathcal{C}_{0}\right)$ say, is still minimal and can be transformed to a system in canonical form $\left(\tilde{\mathcal{A}}_{0}, \tilde{\mathcal{B}}_{0}, \tilde{\mathcal{C}}_{0}\right)$ with fewer unrestricted entries in $\tilde{\mathcal{B}}_{0}$.

Second, if at least one of the row ranks of the p.u.t. blocks decreases in the limit, the limiting system is no longer minimal. Consequently, $\left(\tilde{\Omega}_{S}, \tilde{p}\right)<\left(\Omega_{S}, p\right)$ in the limit.
To illustrate this point consider again Example 4 with Equation (12) rewritten as

$$
x_{t+1,1}=x_{t, 1}+x_{t, 2}+\mathcal{B}_{1,1} \varepsilon_{t}, \quad x_{t+1,2}=x_{t, 2}+\mathcal{B}_{1,2,1} \varepsilon_{t}, \quad x_{t+1,3}=x_{t, 3}+\mathcal{B}_{1,2,2} \varepsilon_{t} .
$$

If $\mathcal{B}_{1,2,1}=\left[0, b_{1,2,1,2}\right] \neq 0$ and $\mathcal{B}_{1,2,2}=\left[b_{1,2,2,1}, b_{1,2,2,2}\right] \neq 0, b_{1,2,2,1}>0$, it holds that $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ is an $\mathrm{I}(2)$ process with state space unit root structure $\Omega_{S}=((0,1,2))$.
Now consider a sequence of systems with all parameters except for $b_{1,2,1,2}$ constant and $b_{1,2,1,2} \rightarrow 0$. The limiting system is then given by

$$
\begin{aligned}
y_{t} & =\mathcal{C}_{1,1}^{E} x_{t, 1}+\mathcal{C}_{1,2}^{G} x_{t, 2}+\mathcal{C}_{1,2}^{E} x_{t, 3}+\varepsilon_{t}, \\
{\left[\begin{array}{l}
x_{t+1,1} \\
x_{t+1,2} \\
x_{t+1,3}
\end{array}\right] } & =\left[\begin{array}{lll}
1 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right]\left[\begin{array}{l}
x_{t, 1} \\
x_{t, 2} \\
x_{t, 3}
\end{array}\right]+\left[\begin{array}{cc}
b_{1,1,1} & b_{1,1,2} \\
0 & 0 \\
b_{1,2,2,1} & b_{1,2,2,2}
\end{array}\right] \varepsilon_{t,} \quad x_{1,1}=x_{1,2}=x_{1,3}=0 .
\end{aligned}
$$

In the limiting system $x_{t, 2}=0$ is redundant and $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ is an $\mathrm{I}(1)$ process rather than an $\mathrm{I}(2)$ process. Dropping $x_{t, 2}$ leads to a state space realisation of the limiting system $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ given by

$$
\begin{aligned}
y_{t} & =\mathcal{C}_{1,1}^{E} x_{t, 1}+\mathcal{C}_{1,2}^{E} x_{t, 3}+\varepsilon_{t}=\tilde{C} \tilde{x}_{t}+\varepsilon_{t}, \quad \tilde{x}_{t} \in \mathbb{R}^{2}, \\
\tilde{x}_{t+1}=\left[\begin{array}{l}
x_{t+1,1} \\
x_{t+1,3}
\end{array}\right] & =\left[\begin{array}{ll}
1 & 0 \\
0 & 1
\end{array}\right]\left[\begin{array}{l}
x_{t, 1} \\
x_{t, 3}
\end{array}\right]+\left[\begin{array}{cc}
b_{1,1,1} & b_{1,1,2} \\
b_{1,2,2,1} & b_{1,2,2,2}
\end{array}\right] \varepsilon_{t}=\tilde{x}_{t}+\tilde{B} \varepsilon_{t}, \quad x_{1,1}=x_{1,3}=0 .
\end{aligned}
$$

In case $\tilde{B}$ has full rank, the above system is minimal. Since $b_{1,2,2,1}>0$, the matrix $\tilde{B}$ needs to be transformed into p.u.t. format. By definition all systems in the sequence, with $b_{1,2,1,2} \neq 0$, have structure indices $p=[0,2,1]^{\prime}$ as discussed in Example 12. The limiting system-in case of full rank of $\tilde{B}$-has indices $\tilde{p}=[1,2]^{\prime}$. To relate to Definition 9 choose the permutation matrix $S=\left[\begin{array}{lll}1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0\end{array}\right]$ to arrive at

$$
S \mathcal{A}_{u} S^{\prime}=\left[\begin{array}{ccc}
1 & 0 & 1 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right]=\left[\begin{array}{cc}
I_{2} & \tilde{J}_{12} \\
0 & \tilde{J}_{2}
\end{array}\right], \quad S p=\left[\begin{array}{l}
0 \\
1 \\
2
\end{array}\right]=\left[\begin{array}{c}
\left(p_{1}\right)_{1} \\
\left(p_{1}\right)_{2} \\
p_{2}
\end{array}\right]
$$

This shows that $(\tilde{p})_{i}>\left(p_{1}\right)_{i}, i=1,2$ and thus the limiting system has a smaller multi-index $\Gamma$ than the systems of the sequence. In case $\tilde{B}$ has reduced rank equal to one a further reduction in the system order to $n=1$ along similar lines as discussed is possible, again leading to a limiting system with smaller multi-index $\Gamma$.

The discussion shows that the closure of $\Delta_{\Omega_{S}, p}^{\mathcal{B}}$ is related to lower order systems in the sense of Definition 9. The precise statement is given in Theorem 3 after a discussion of the closure of the stable subsystems.

### 4.1.2. The Closure of $\Delta_{\alpha}$

Consider a convergent sequence of systems $\left\{\left(\mathcal{A}_{j}, \mathcal{B}_{j}, \mathcal{C}_{j}\right)\right\}_{j \in \mathbb{N}}$ in $\Delta_{\alpha_{\bullet}}$ and denote the limiting system by $\left(A_{0}, B_{0}, C_{0}\right)$. Clearly, $\lambda_{|\max |}\left(A_{0}\right) \leq 1$ holds true for the limit $A_{0}$ of the sequence $\left\{\mathcal{A}_{j}\right\}_{j \in \mathbb{N}}$ with $\lambda_{|\max |}\left(\mathcal{A}_{j}\right)<1$ for all $j$. Therefore, two cases have to be discussed for the limit:

- If $\lambda_{|\max |}\left(A_{0}\right)<1$, the potentially non-minimal limiting system $\left(A_{0}, B_{0}, C_{0}\right)$ corresponds to a minimal state space realization with Kronecker indices smaller or equal to $\alpha_{\bullet}$ (cf. Hannan and Deistler 1988, Theorem 2.5.3).
- If $\lambda_{|\max |}\left(A_{0}\right)=1$, the limiting matrix $A_{0}$ is similar to a block matrix $\tilde{A}=\operatorname{diag}\left(\tilde{J}_{2}, \tilde{A}_{\bullet}\right)$, where all eigenvalues of $\tilde{J}_{2}$ have unit modulus and $\lambda_{|\max |}\left(\tilde{A}_{\bullet}\right)<1$.

The first case is well understood, compare Hannan and Deistler (1988, chp. 2), since the limit in this case corresponds to a stable transfer function. In the second case the limiting system can be separated into two subsystems $\left(\tilde{J}_{2}, \tilde{B}_{u}, \tilde{C}_{u}\right)$ and $\left(\tilde{A}_{\bullet}, \tilde{B}_{\bullet}, \tilde{C}_{\bullet}\right)$, according to the block diagonal structure of $\tilde{A}$. The state space unit root structure of the limiting system $\left(A_{0}, B_{0}, C_{0}\right)$ depends on the multiplicities of the eigenvalues of the matrix $\tilde{J}_{2}$ and is greater (in the sense of Definition 9) than the empty state space unit root structure. At the same time the Kronecker indices of the subsystem $\left(\tilde{A}_{\bullet}, \tilde{B}_{\bullet}, \tilde{C}_{\bullet}\right)$ are smaller than $\alpha_{\bullet}$, compare again Hannan and Deistler (1988, chp. 2). Since the Kronecker indices impose restrictions on some entries of the matrices $\mathcal{A}_{j}$ and thus also on $A_{0}$, the block $\tilde{J}_{2}$ and consequently also the limiting state space unit root structure might be subject to further restrictions.

### 4.1.3. The Conformable Index Set and the Closure of $\Delta_{\Gamma}$

The previous subsection shows that the closure of $\Delta_{\Gamma}$ does not only contain systems corresponding to transfer functions with multi-index smaller or equal to $\Gamma$, but also systems that are related in a different way that is formalized below.

Definition 10 (Conformable index set). Given a multi-index $\Gamma=\left(\Omega_{S}, p, \alpha_{\bullet}\right)$, the set of conformable multi-indices $\mathcal{K}(\Gamma)$ contains all multi-indices $\tilde{\Gamma}=\left(\tilde{\Omega}_{S}, \tilde{p}, \tilde{\chi}_{\bullet}\right)$, where:

- The pair $\left(\tilde{\Omega}_{S}, \tilde{p}\right)$ with corresponding matrix $\tilde{\mathcal{A}}_{u}$ in canonical form extends $\left(\Omega_{S}, p\right)$ with corresponding matrix $\mathcal{A}_{u}$ in canonical form, i.e., there exists a permutation matrix $S$ such that

$$
S \tilde{\mathcal{A}}_{u} S^{\prime}=\left[\begin{array}{cc}
\mathcal{A}_{u} & 0 \\
0 & \tilde{J}_{2}
\end{array}\right] \quad \text { and } \quad S \tilde{p}=\left[\begin{array}{c}
p \\
\tilde{p}_{2}
\end{array}\right]
$$

- $\tilde{\alpha}_{\bullet} \leq \alpha_{\bullet}$.
- $\quad \tilde{n}_{u}+\tilde{n}_{\bullet}=n_{u}+n_{\bullet}$.

Please note that the definition implies $\Gamma \in \mathcal{K}(\Gamma)$. The importance of the set $\mathcal{K}(\Gamma)$ is clarified in the following theorem:

Theorem 3. Transfer functions corresponding to state space realizations with multi-index $\tilde{\Gamma} \leq \Gamma$ are contained in the set $\pi\left(\overline{\Delta_{\Gamma}}\right)$. The set $\pi\left(\overline{\Delta_{\Gamma}}\right)$ is contained in the union of all sets $M_{\Gamma}$ for $\check{\Gamma} \leq \tilde{\Gamma}$ with $\tilde{\Gamma}$ conformable to $\Gamma$, i.e.,

$$
\bigcup_{\tilde{\Gamma} \leq \Gamma} M_{\tilde{\Gamma}} \subset \pi\left(\overline{\Delta_{\Gamma}}\right) \subset \bigcup_{\tilde{\Gamma} \in \mathcal{K}(\Gamma)} \bigcup_{\check{\Gamma} \leq \tilde{\Gamma}} M_{\check{\Gamma}} .
$$

Theorem 3 provides a characterization of the transfer functions corresponding to systems in the closure of $\Delta_{\Gamma}$. The conformable set $\mathcal{K}(\Gamma)$ plays a key role here, since it characterizes the set of all minimal systems that can be obtained as limits of convergent sequences from within the set $\Delta_{\Gamma}$. Conformable indices extend the matrix $\mathcal{A}_{u}$ corresponding to the unit root structure by the block $\tilde{J}_{2}$.

The second inclusion in Theorem 3 is potentially strict, depending on the Kronecker indices $\alpha_{\bullet}$ in $\Gamma$. Equality holds, e.g., in the following case:

Corollary 3. For every multi-index $\Gamma$ with $n_{\bullet}=0$ the set of conformable indices consists only of $\Gamma$, which implies $\pi\left(\overline{\Delta_{\Gamma}}\right)=\bigcup_{\tilde{\Gamma} \leq \Gamma} M_{\tilde{\Gamma}}$.

### 4.2. The Closure of $M_{\Gamma}$

It remains to investigate the closure of $M_{\Gamma}$ in $M$. Hannan and Deistler (1988, Theorem 2.6 .5 (ii) and Remark 3, p. 73) show that for any order $n$, there exist Kronecker indices $\alpha_{\bullet, g}=\alpha_{\bullet, g}(n)$ corresponding to the generic neighborhood $M_{\alpha_{\bullet, g}}$ for transfer functions of order $n$ such that

$$
M_{\bullet}, n:=\bigcup_{\alpha_{\bullet} \mid n_{\bullet}\left(\alpha_{\bullet}\right)=n} M_{\alpha_{\bullet}} \subset \overline{M_{\alpha_{\bullet}, g}}
$$

where $M_{\alpha_{\bullet}}:=\pi\left(\Delta_{\alpha_{\bullet}}\right)$. Here $M_{\bullet, n}$ denotes the set of all transfer functions of order $n$ with state space realizations $(A, B, C)$ satisfying $\lambda_{|\max |}(A)<1$. Every transfer function in $M_{\bullet}, n$ can be approximated by a sequence of transfer functions in $M_{\alpha_{\bullet, g}}$.

It can be easily seen that a generic neighborhood also exists for systems with state space unit root structure $\Omega_{S}$ and without stable subsystem: Set the structure indices $p$ to have a minimal number of elements restricted in p.u.t. sub-blocks of $\mathcal{B}_{u}$, i.e., for any block $\mathcal{B}_{k, h_{k}, j} \in \mathbb{C}^{n_{k, h_{k}, j} \times s}$, or $\mathcal{B}_{k, h_{k}, j} \in \mathbb{R}^{n_{k, h_{k}, j} \times s}$ in case of a real unit root, set the corresponding structure indices to $p=\left[1, \ldots, n_{k, h_{k}, j}\right]$. Any p.u.t. matrix can be approximated by a matrix in this generic neighborhood with some positive entries restricted by the p.u.t. structure tending to zero. Combining these results with Theorem 3 implies the existence of a generic neighborhood for the canonical form considered in this paper:

Theorem 4. Let $M\left(\Omega_{S}, n_{\bullet}\right)$ be the set of all transfer functions $k(z) \in M_{n_{u}\left(\Omega_{S}\right)+n_{\bullet}}$ with state space unit root structure $\Omega_{S}$. For every $\Omega_{S}$ and $n_{\bullet}$, there exists a multi-index $\Gamma_{g}:=\Gamma_{g}\left(\Omega_{S}, n_{\bullet}\right)$ such that

$$
\begin{equation*}
M\left(\Omega_{S}, n_{\bullet}\right) \subset \overline{M_{\Gamma_{g}}} . \tag{14}
\end{equation*}
$$

Moreover, it holds that $M\left(\Omega_{S}, n_{\bullet}\right) \subset \overline{M_{\alpha_{\bullet, g}(n)}}$ for every $\Omega_{S}$ and $n_{\bullet}$ satisfying $n_{u}\left(\Omega_{S}\right)+n_{\bullet} \leq n$.
Theorem 4 is the basis for choosing a generic multi-index $\Gamma$ for maximizing the pseudo likelihood function. For every $\Omega_{S}$ and $n_{\bullet}$ there exists a generic piece that-in its closure-contains all transfer functions of order $n_{u}\left(\Omega_{S}\right)+n_{\bullet}$ and state space unit root structure $\Omega_{S}$ : The set of transfer functions corresponding to the multi-index with the largest possible structure indices $p$ in the sense of Definition 9 (iii) and generic Kronecker indices for the stable subsystem. Choosing these sets and their corresponding parameter spaces as model sets is, therefore, the most convenient choice for numerical maximization, if only $\Omega_{S}$ and $n_{\bullet}$ are known.

If, e.g., only an upper bound for the system order $n$ is known and the goal is only to obtain consistent estimators, using $\alpha_{\bullet, g}(n)$ is a feasible choice, since all transfer functions in the closure of the set $M_{\alpha_{0, g}(n)}$ can be approximated arbitrarily well, regardless of their potential state space unit root structure $\Omega_{S}, n_{u}\left(\Omega_{S}\right) \leq n$. For testing hypotheses, however, it is important to understand the topological relations between sets corresponding to different multi-indices $\Gamma$. In the following we focus on the multi-indices $\Gamma_{g}\left(\Omega_{S}, n_{\bullet}\right)$ for arbitrary $\Omega_{S}$ and $n_{\bullet}$.

The closure of $M\left(\Omega_{S}, n_{\bullet}\right)$ contains also transfer functions that have a different state space unit root structure than $\Omega_{S}$. Considering convergent sequences of state space realizations $\left(A_{j}, B_{j}, C_{j}\right)_{j \in \mathbb{N}}$ of transfer functions in $M\left(\Omega_{S}, n_{\bullet}\right)$, the state space unit root structure of $\left(A_{0}, B_{0}, C_{0}\right):=\lim _{j \rightarrow \infty}\left(A_{j}, B_{j}, C_{j}\right)$ may differ in three ways:

- For sequences $\left(\mathcal{A}_{j}, \mathcal{B}_{j}, \mathcal{C}_{j}\right)_{j \in \mathbb{N}}$ in canonical form rows of $\mathcal{B}_{u, j}$ can tend to zero, which reduces the state space unit root structure as discussed in Section 4.1.1.
- Stable eigenvalues of $A_{j}$ may converge to the unit circle, thereby extending the unit root structure.
- Off-diagonal entries of the sub-block $\mathcal{A}_{u, j}$ of $\mathcal{A}_{j}=T_{j} A_{j} T_{j}^{-1}$ may be converging to zeros in the sub-block $\mathcal{A}_{u, 0}$ of the limit $\mathcal{A}_{0}=T_{0} A_{0} T_{0}^{-1}$ in canonical form, resulting in a different attainable state space unit root structure. Here $T_{j} \in \mathbb{C}^{n \times n}$ for all $j \in \mathbb{N}$ are regular matrices transforming $A_{j}$ to canonical form and $T_{0} \in \mathbb{C}^{n \times n}$ transforms $A_{0}$ accordingly.
The first change of $\Omega_{S}$ described above results in a transfer function with smaller state space unit root structure according to Definition 9 (ii). The implications of the other two cases are summarized in the following definition:

Definition 11 (Attainable unit root structures). For given $n_{\bullet}$ and $\Omega_{S}$ the set $\mathcal{A}\left(\Omega_{S}, n_{\bullet}\right)$ of attainable unit root structures contains all pairs $\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right)$, where $\tilde{\Omega}_{S}$ with corresponding matrix $\tilde{\mathcal{A}}_{u}$ in canonical form extends $\Omega_{S}$ with corresponding matrix $\mathcal{A}_{u}$ in canonical form, i.e., there exists a permutation matrix $S$ such that

$$
S \tilde{\mathcal{A}}_{u} S^{\prime}=\left[\begin{array}{cc}
\check{\mathcal{A}}_{u} & J_{12} \\
0 & J_{2}
\end{array}\right]
$$

where $\check{\mathcal{A}}_{u}$ can be obtained by replacing off-diagonal entries in $\mathcal{A}_{u}$ by zeros and where $\tilde{n}_{\bullet}:=n_{\bullet}-d_{J}$ with $d_{J}$ the dimension of $J_{2} \in \mathbb{C}^{d_{J} \times d_{J}}$.

Remark 17. It is a direct consequence of the definition of $\mathcal{A}\left(\Omega_{S}, n_{\bullet}\right)$ that $\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right) \in \mathcal{A}\left(\Omega_{S}, n_{\bullet}\right)$ implies $\mathcal{A}\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right) \subset \mathcal{A}\left(\Omega_{S}, n_{\bullet}\right)$.

## Theorem 5.

(i) $M_{\Gamma}$ is $T_{p t}$-open in $\overline{M_{\Gamma}}$ (see Definition 8 for a definition of $T_{p t}$ ).
(ii) For every generic multi-index $\Gamma_{g}$ corresponding to $\Omega_{S}$ and $n$. it holds that

$$
\begin{aligned}
\pi\left(\overline{\Delta_{\Gamma_{g}}}\right) & \subset \bigcup_{\tilde{\Gamma} \in \mathcal{K}\left(\Gamma_{g}\right)} \bigcup_{\check{\Gamma} \leq \tilde{\Gamma}} M_{\check{\Gamma}} \\
& \subset \bigcup_{\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right) \in \mathcal{A}\left(\Omega_{S}, n_{\bullet}\right)\left(\check{\Omega}_{S}, \check{n}_{\bullet}\right) \leq\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right)} M\left(\check{\Omega}_{S}, \check{n}_{\bullet}\right)
\end{aligned}
$$

Theorem 5 has important consequences for statistical analysis, e.g., PML estimation, since-as stated several times already-maximizing the pseudo likelihood function over $\Theta_{\Gamma}$ effectively amounts to calculating the supremum over the larger set $\overline{M_{\Gamma}}$. Depending on the choice of $\Gamma$ the following asymptotic behavior may occur:

- If $\Gamma$ is chosen correctly and the estimator of the transfer function is consistent, openness of $M_{\Gamma}$ in its closure implies that the probability of the estimator being an interior point of $M_{\Gamma}$ tends to one asymptotically. Since the mapping attaching the parameters to the transfer function is continuous on an open and dense set, consistency in terms of transfer functions, therefore, implies generic consistency of the parameter estimators.
- If the multi-index is incorrectly chosen to equal $\Gamma$, estimator consistency is still possible if the true multi-index $\Gamma_{0}<\Gamma$, as in this case $M_{\Gamma_{0}} \subset \bar{M}_{\Gamma}$. This is in some sense not too surprising and something that is also well-known in the simpler VAR framework where consistency of OLS can be established when the true autoregressive order is smaller than the order chosen for estimation. Analogous to the lag number in the VAR case, thus, a necessary condition for consistency is to choose the system order larger or equal to the true system order.

Finally, note that Theorem 5 also implies the following result relevant for the determination of the unit root structure, further discussed in Sections 5.1.1 and 5.2.1:

Corollary 4. For every pair $\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right) \in \mathcal{A}\left(\Omega_{S}, n_{\bullet}\right)$ it holds that

$$
\overline{M\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right)} \subset \overline{M\left(\Omega_{S}, n_{\bullet}\right)}
$$

## 5. Testing Commonly Used Hypotheses in the MFI(1) and I(2) Cases

This section discusses a large number of hypotheses, respectively restrictions, on cointegrating spaces, adjustment coefficients and deterministic components often tested in the empirical literature. As with the VECM framework, as discussed for the I(2) case in Section 2, testing hypotheses on the cointegrating spaces or adjustment coefficients may necessitate different reparameterizations.

### 5.1. The MFI (1) Case

The two by far most widely used cases of $\mathrm{MFI}(1)$ processes are $I(1)$ processes and seasonally (co-)integrated processes for quarterly data with state space unit root structure $\left(\left(0, d_{1}^{1}\right),\left(\pi / 2, d_{1}^{2}\right),\left(\pi, d_{1}^{3}\right)\right)$. In general, assuming for notational simplicity $\omega_{1}=0$ and $\omega_{l}=\pi$, it holds that for $t>0$ and $x_{1, u}=0$ we have

$$
\begin{aligned}
y_{t}= & \sum_{k=1}^{l} \mathcal{C}_{k, \mathbb{R}} x_{t, k, \mathbb{R}}+\mathcal{C}_{\bullet} x_{t, \bullet}+\Phi d_{t}+\varepsilon_{t} \\
= & \mathcal{C}_{1} x_{t, 1}+\sum_{k=2}^{l-1}\left(\mathcal{C}_{k} x_{t, k}+\overline{\mathcal{C}}_{k} \bar{x}_{t, k}\right)+\mathcal{C}_{l} x_{t, l}^{j}+\mathcal{C}_{\bullet} x_{t, \bullet}+\Phi d_{t}+\varepsilon_{t} \\
= & \mathcal{C}_{1} \mathcal{B}_{1} \sum_{j=1}^{t-1} \varepsilon_{t-j}+2 \sum_{k=2}^{l-1} \mathcal{R}\left(\mathcal{C}_{k} \mathcal{B}_{k}^{t-1} \sum_{j=1}^{t-1}\left(\bar{z}_{k}\right)^{j-1} \varepsilon_{t-j}\right)+\mathcal{C}_{l} \mathcal{B}_{l}^{t-1} \sum_{j=1}^{t}(-1)^{j-1} \varepsilon_{t-j} \\
& +\mathcal{C} \cdot \sum_{j=1}^{t-1} \mathcal{A}_{\bullet}^{j-1} \mathcal{B}_{\bullet} \varepsilon_{t-j}+\mathcal{C}_{\bullet} \mathcal{A}_{\bullet}^{t-1} x_{1, \bullet}+\Phi d_{t}+\varepsilon_{t} \\
= & \mathcal{C}_{1} \mathcal{B}_{1} \sum_{j=1}^{t-1} \varepsilon_{t-j}+2 \sum_{k=2}^{l-1} \sum_{j=1}^{t-1}\left(\mathcal{R}\left(\mathcal{C}_{k} \mathcal{B}_{k}\right) \cos \left(\omega_{k}(j-1)\right)+\mathcal{I}\left(\mathcal{C}_{k} \mathcal{B}_{k}\right) \sin \left(\omega_{k}(j-1)\right)\right) \varepsilon_{t-j} \\
& +\mathcal{C}_{l} \mathcal{B}_{l} \sum_{j=1}^{t-1}(-1)^{j-1} \varepsilon_{t-j}+\mathcal{C}_{\bullet}^{t-1} \sum_{j=1}^{t} \mathcal{A}_{\bullet}^{j-1} \mathcal{B}_{\bullet} \varepsilon_{t-j}+\mathcal{C}_{\bullet} \mathcal{A}_{\bullet}^{t-1} x_{1, \bullet}+\Phi d_{t}+\varepsilon_{t} .
\end{aligned}
$$

The above equation provides an additive decomposition of $\left\{y_{t}\right\}_{t \in \mathbb{Z}}$ into stochastic trends and cycles, the deterministic and stationary components. The stochastic cycles at frequency $0<\omega_{k}<\pi$ are, of course, given by the combination of sine and cosine terms. For the MFI(1) case this can also be seen directly from considering the real valued canonical form discussed in Remark 4, with the matrices $\mathcal{A}_{k, \mathbb{R}}$ for $k=2, \ldots, l-1$, given by $\mathcal{A}_{k, \mathbb{R}}=I_{d_{1}^{k}} \otimes\left(\begin{array}{cc}\cos \left(\omega_{k}\right) & -\sin \left(\omega_{k}\right) \\ \sin \left(\omega_{k}\right) & \cos \left(\omega_{k}\right)\end{array}\right)$ in this case.

The ranks of $\mathcal{C}_{k} \mathcal{B}_{k}$ are equal to the integers $d_{1}^{k}$ in $\Omega_{S}=\left(\left(\omega_{1}, d_{1}^{1}\right), \ldots,\left(\omega_{l}, d_{1}^{l}\right)\right)$. The number of stochastic trends is equal to $d_{1}^{1}$, the number of stochastic cycles at frequency $\omega_{k}$ is equal to $2 d_{1}^{k}$ for $k=2, \ldots, l-1$ and equal to $d_{1}^{l}$ if $k=l$, as discussed in Section 3.

Moreover, in the $\operatorname{MFI}(1)$ case, $d_{1}^{k}$ is linked to the complex cointegrating rank $r_{k}$ at frequency $\omega_{k}$, defined in Johansen (1991) and Johansen and Schaumburg (1999) in the VECM case as the rank of the matrix $\Pi_{k}:=-a\left(z_{k}\right)$. For VARMA processes with arbitrary integration orders the complex cointegrating rank $r_{k}$ at frequency $\omega_{k}$ is $r_{k}:=\operatorname{rank}\left(-k^{-1}\left(z_{k}\right)\right)$, where $k(z)$ is the transfer function, with $r_{k}=s-d_{1}^{k}$ in the $\operatorname{MFI}(1)$ case. Thus, in the $\operatorname{MFI}(1)$ case, determination of the state space unit root structure corresponds to determination of the complex cointegrating ranks in the VECM case.

In the VECM setting, the matrix $\Pi_{k}$ is usually factorized into $\Pi_{k}=\alpha_{k} \beta_{k}^{\prime}$, as presented for the $\mathrm{I}(1)$ case in Section 2. For $\omega_{k}=\{0, \pi\}$ the column space of $\beta_{k}$ gives the cointegrating space of the process at frequency $\omega_{k}$. For $0<\omega_{k}<\pi$ the relation between the column space of $\beta_{k}$ and the space of CIVs and PCIVs at the corresponding frequency is more involved. The columns of $\beta_{k}$ are orthogonal to the columns of $\mathcal{C}_{k}$, the sub-block of $\mathcal{C}$ from a state space realization $(\mathcal{A}, \mathcal{B}, \mathcal{C})$ in canonical form corresponding to the VAR process. Analogously, the column space of the matrix $\alpha_{k}$, containing the so-called adjustment coefficients, is orthogonal to the row space of the sub-block $\mathcal{B}_{k}$ of $\mathcal{B}$.

Both integers $d_{1}^{k}$ and $r_{k}$ are related to the dimensions of the static and dynamic cointegrating spaces in the MFI(1) case: For $\omega_{k} \in\{0, \pi\}$, the cointegrating rank $r_{k}=s-d_{1}^{k}$ coincides with the dimension of the static cointegrating space at frequency $\omega_{k}$. Furthermore, the dimension of the static cointegrating space at frequency $0<\omega_{k}<\pi$ is bounded from above by $r_{k}=s-d_{1}^{k}$, since it is spanned by at most $s-d_{1}^{k}$ vectors $\beta \in \mathbb{R}^{s}$ orthogonal to the complex valued matrix $\mathcal{C}_{k}$. The dimension of the dynamic cointegrating space at $0<\omega_{k}<\pi$ is equal to $2 r_{k}=2\left(s-d_{1}^{k}\right)$. Identifying again $\beta(z)=\beta_{0}+\beta_{1} z$ with the vector $\left[\beta_{0}^{\prime}, \beta_{1}^{\prime}\right]^{\prime}$, a basis of the dynamic cointegrating space at $0<\omega_{k}<\pi$ is then given by the column space of the product

$$
\left[\begin{array}{cc}
\gamma_{0} & \tilde{\gamma}_{0} \\
\gamma_{1} & \tilde{\gamma}_{1}
\end{array}\right]:=\left[\begin{array}{cc}
I_{s} & 0_{s \times s} \\
-\cos \left(\omega_{k}\right) I_{s} & \sin \left(\omega_{k}\right) I_{s}
\end{array}\right]\left[\begin{array}{cc}
\mathcal{R}\left(\beta_{k}\right) & \mathcal{I}\left(\beta_{k}\right) \\
-\mathcal{I}\left(\beta_{k}\right) & \mathcal{R}\left(\beta_{k}\right)
\end{array}\right]
$$

with the columns of $\beta_{k} \in \mathbb{C}^{s \times\left(s-d_{1}^{k}\right)}$ spanning the orthogonal complement of the column space of $\mathcal{C}_{k}$, i.e., $\beta_{k}$ is of full rank and $\beta_{k}^{\prime} \mathcal{C}_{k}=\left(\mathcal{R}\left(\beta_{k}\right)^{\prime}-i \mathcal{I}\left(\beta_{k}\right)^{\prime}\right) \mathcal{C}_{k}=0$. This holds true, since both factors are of full rank and $\left[\gamma_{0}^{\prime}, \gamma_{1}^{\prime}\right]^{\prime}$ satisfies $\left(\bar{z}_{k} \gamma_{0}^{\prime}+\gamma_{1}^{\prime}\right) \mathcal{C}_{k}=0$, which corresponds to the necessary condition given in Example 2 for the columns of $\left[\gamma_{0}^{\prime}, \gamma_{1}^{\prime}\right]^{\prime}$ to be PCIVs. The latter implies $\left(\bar{z}_{k} \tilde{\gamma}_{0}^{\prime}+\tilde{\gamma}_{1}^{\prime}\right) \mathcal{C}_{k}=0$ also for $\left[\tilde{\gamma}_{0}^{\prime}, \tilde{\gamma}_{1}^{\prime}\right]^{\prime}$, highlighting again the additional structure of the cointegrating space emanating from the complex conjugate pairs or eigenvalues (and matrices) as discussed in Example 2.

Please note that the relations between $r_{k}$ and $d_{1}^{k}$ discussed above only hold in the $\operatorname{MFI}(1)$ and $\mathrm{I}(1)$ special cases. For higher orders of integration no such simple relations exist.

In the MFI(1) setting the deterministic component typically includes a constant, seasonal dummies and a linear trend. As discussed in Remark 6, a sufficiently rich set of deterministic components allows to absorb non-zero initial values $x_{1, u}$.

### 5.1.1. Testing Hypotheses on the State Space Unit Root Structure

Using the generic sets of transfer functions $M_{\Gamma_{g}}$ presented in Theorem 4, we can construct pseudo likelihood ratio tests for different hypotheses $H_{0}:\left(\Omega_{S}, n_{\bullet}\right)=\left(\Omega_{S, 0}, n_{\bullet}, 0\right)$ against chosen alternatives. Note, however, that by the results of Theorem 5 the null hypothesis includes all pairs $\left(\Omega_{S}, n_{\bullet}\right) \in$ $\mathcal{A}\left(\Omega_{S, 0}, n_{\bullet}, 0\right)$ as well as all pairs $\left(\Omega_{S}, n_{\bullet}\right)$ that are smaller than a pair $\left(\Omega_{S}, \tilde{n}_{\bullet}\right) \in \mathcal{A}\left(\Omega_{S, 0}, n_{\bullet}, 0\right)$.

As common in the VECM setting, first consider hypotheses at a single frequency $\omega_{k}$. For an $\operatorname{MFI}(1)$ process, the hypothesis of a state space unit root structure equal to $\Omega_{S, 0}=\left(\left(\omega_{k}, d_{1,0}^{k}\right)\right)$ corresponds to the hypothesis of the (compex) cointegrating rank $r_{k}$ at frequency $\omega_{k}$ being equal to $r_{0}=s-d_{1,0}^{k}$. Maximization of the pseudo likelihood function over the set $\overline{M\left(\left(\left(\omega_{k}, d_{1,0}^{k}\right)\right), n-\delta_{k} d_{1,0}^{k}\right)}-$ with a suitably chosen order $n$-leads to estimates that may be arbitrary close to transfer functions with different state space unit root structures $\Omega_{S}$. These include $\Omega_{S}$ with additional unit root frequencies $\omega_{\tilde{k}}$, with the integers $d_{1}^{\tilde{k}}$ restricted only by the order $n$. Therefore, focusing on a single frequency $\omega_{k}$ does not rule out a more complicated true state space unit root structure. Assume $n \geq \delta_{k} s$ with $\delta_{k}=1$ for $\omega_{k} \in\{0, \pi\}$ and $\delta_{k}=2$ else. Corollary 4 shows that

$$
\overline{M(\}, n)} \supset \overline{M\left(\left(\left(\omega_{k}, 1\right)\right), n-\delta_{k}\right)} \supset \cdots \supset \overline{M\left(\left(\left(\omega_{k}, s\right)\right), n-s \delta_{k}\right)}
$$

since, e.g., $\left(\left(\left(\omega_{k}, 1\right)\right), n-\delta_{k}\right) \in \mathcal{A}(\}, n)$.
Analogously to the procedure of testing for the complex cointegrating rank $r_{k}$ in the VECM setting, these inclusions can be employed to test for $d_{1}^{k}$ : Start with the hypothesis of $d_{1}^{k}=s$ against the alternative of $0 \leq d_{1}^{k}<s$ and decrease the assumed $d_{1}^{k}$ consecutively until the test does not reject the null hypothesis.

Furthermore, one can formulate hypotheses on $d_{1}^{k}$ jointly at different frequencies $\omega_{k}$. Again, there exist inclusions based on the definition of the set of attainable state space unit root structures and Corollary 4, which can be used to consecutively test hypotheses on $\Omega_{S}$.

### 5.1.2. Testing Hypotheses on CIVs and PCIVs

Johansen (1995) considers in the $I(1)$ case three types of hypotheses on the cointegrating space spanned by the columns of $\beta$ that are each motivated by examples from economic research: The different cases correspond to different types of hypotheses related to restrictions implied by economic theory.
(i) $H_{0}: \beta=H \varphi, \beta \in \mathbb{R}^{s \times r}, H \in \mathbb{R}^{s \times t}, \varphi \in \mathbb{R}^{t \times r}, r \leq t<s$ : The cointegrating space is known to be a subspace of the column space of $H$ (which is of full column rank).
(ii) $H_{0}^{\prime}: \beta=[b, \varphi], \beta \in \mathbb{R}^{s \times r}, b \in \mathbb{R}^{s \times t}, \varphi \in \mathbb{R}^{s \times r-t}, 0<t \leq r$ : Some cointegrating relations are known.
(iii) $H_{0}^{\prime \prime}: \beta=\left[H_{1} \varphi_{1}, \ldots, H_{c} \varphi_{c}\right], \beta \in \mathbb{R}^{s \times r}, H_{j} \in \mathbb{R}^{s \times t_{j}}, \varphi_{j} \in \mathbb{R}^{t_{j} \times r_{j}}, r_{j} \leq t_{j} \leq s$, for $j=1, \ldots, c$ such that $\sum_{j=1}^{c} r_{j}=r$. Cointegrating relations are known to be in the column spaces of matrices $H_{k}$ (which are of full column rank).

As discussed in Example 1, cointegration at $\omega_{k}=0$ occurs if and only if a vector $\beta_{j}$ satisfies $\beta_{j}^{\prime} \mathcal{C}_{1}=0$. In other words, the column space of $\mathcal{C}_{1}$ is the orthocomplement of the cointegrating space spanned by the columns of $\beta$ and hypotheses on $\beta$ restrict entries of $\mathcal{C}_{1}$.

The first type of hypothesis, $H_{0}$, implies that the column space of $\mathcal{C}_{1}$ is equal to the orthocomplement of the column space of $H \varphi$. Assume w.l.o.g. $H \in O_{s, t}, \varphi_{\perp} \in O_{t, t-r}$ and $H_{\perp} \in O_{s, s-t}$, such that the columns of $\left[H \varphi_{\perp}, H_{\perp}\right]$ form an orthonormal basis for the orthocomplement of the cointegrating space. Consider now the mapping:

$$
\mathcal{C}_{1}^{r}\left(\check{\boldsymbol{\theta}}_{L}, \boldsymbol{\theta}_{R}\right):=\left[H \cdot \check{R}_{L}\left(\check{\boldsymbol{\theta}}_{L}\right)^{\prime}\left[\begin{array}{c}
I_{t-r}  \tag{15}\\
0_{r \times(t-r)}
\end{array}\right], \quad H_{\perp}\right] \cdot R_{R}\left(\boldsymbol{\theta}_{R}\right),
$$

where $\check{R}_{L}\left(\check{\theta}_{L}\right):=\prod_{i=1}^{t-r} \prod_{j=1}^{r} R_{t, i, t-r+j}\left(\theta_{L, r(i-1)+j}\right) \in \mathbb{R}^{t \times t}$ and $R_{R}\left(\boldsymbol{\theta}_{R}\right) \in \mathbb{R}^{(s-r) \times(s-r)}$ as in Lemma 1 . From this one can derive a parameterization of the set of matrices $\mathcal{C}_{1}^{r}$ corresponding to $H_{0}$, analogously to Lemma 1. The difference of the number of free parameters under the null hypothesis and under the alternative is the difference between the number of free parameters in $\boldsymbol{\theta}_{L} \in[0,2 \pi)^{r(s-r)}$ and $\check{\theta}_{L} \in[0,2 \pi)^{r(t-r)}$, implying a reduction of the number of free parameters of $r(s-t)$ under the null hypothesis. This necessarily coincides with the number of degrees of freedom of the corresponding test statistic in the VECM setting (cf. Johansen 1995, Theorem 7.2).

The second type of hypothesis, $H_{0}^{\prime}$, is also straightforwardly parameterized: In this case a subspace of the cointegrating space is known and given by the column space of $b \in \mathbb{R}^{s \times t}$. Assume w.l.o.g. $b \in O_{s, t}$. The orthocomplement of $\beta=[b, \varphi]$ is given by the set of matrices $\mathcal{C}_{1}$ satisfying the restriction $b^{\prime} \mathcal{C}_{1}=0$, i.e., the set $O_{s, d_{1}}(b)$ defined in (13). The parameterization of this set has already been discussed. The reduction of the number of free parameters under the null hypothesis is $t(s-r)$ which again coincides with the number of degrees of freedom of the corresponding test statistic in the VECM setting (cf. Johansen 1995, Theorem 7.3).

Finally, the third type of hypothesis, $H_{0}^{\prime \prime}$, is the most difficult to parameterize in our setting. As an illustrative example consider the case $H_{0}^{\prime \prime}: \beta=\left[H_{1} \varphi_{1}, H_{2} \varphi_{2}\right], \beta \in \mathbb{R}^{s \times r}, H_{1} \in \mathbb{R}^{s \times t_{1}}, H_{2} \in \mathbb{R}^{s \times t_{2}}, \varphi_{1} \in$ $\mathbb{R}^{t_{1} \times r_{1}}, \varphi_{2} \in \mathbb{R}^{t_{2} \times r_{2}}, r_{j} \leq t_{j} \leq s$ and $r_{1}+r_{2}=r$. W.l.o.g. choose $H_{b} \in O_{s, t_{b}}$ such that its columns span the $t_{b}$-dimensional intersection of the column spaces of $H_{1}$ and $H_{2}$ and choose $\tilde{H}_{j} \in O_{s, \tilde{t}_{j}}\left(H_{b}\right), j=1,2$ such that the columns of $\tilde{H}_{j}$ and $H_{b}$ span the column space of $H_{j}$. Define $\tilde{H}:=\left[\tilde{H}_{1}, \tilde{H}_{2}, H_{b}\right] \in O_{s, \tilde{t}}$, with $\tilde{t}=\tilde{t}_{1}+\tilde{t}_{2}+t_{b}$. Let w.l.o.g. $\tilde{H}_{\perp} \in O_{s, s-\tilde{t}}(\tilde{H})$ and define $p_{j}:=\min \left(r_{j}, \tilde{t}_{j}\right), q_{j}:=\max \left(r_{j}, \tilde{t}_{j}\right)$ for $j=1,2$ and $p_{b}=q_{1}-\tilde{t}_{1}+q_{2}-\tilde{t}_{2}$. A parameterization of $\beta^{r} \in O_{s, r}$ satisfying the restrictions under the null hypothesis can be derived from the following mapping:

$$
\beta^{r}\left(\boldsymbol{\theta}_{H}, \boldsymbol{\theta}_{R, \beta}\right):=\tilde{H} \cdot R_{H}\left(\boldsymbol{\theta}_{H}\right)^{\prime}\left[\begin{array}{ccc}
I_{p_{1}} & 0_{p_{1} \times p_{2}} & 0_{p_{1} \times p_{b}} \\
\mathbf{0}_{\left(q_{1}-r_{1}\right) \times p_{1}} & 0_{\left(q_{1}-r_{1}\right) \times p_{2}} & 0_{\left(q_{1}-r_{1}\right) \times p_{b}} \\
0_{p_{2} \times p_{1}} & I_{p_{2}} & 0_{p_{2} \times p_{b}} \\
0_{\left(q_{2}-r_{2}\right) \times p_{1}} & \mathbf{0}_{\left(q_{2}-r_{2}\right) \times p_{2}} & 0_{\left(q_{2}-r_{2}\right) \times p_{b}} \\
0_{p_{b} \times p_{1}} & 0_{p_{b} \times p_{2}} & I_{p_{b}} \\
\mathbf{0}_{\left(\tilde{t}-q_{1}-q_{2}\right) \times p_{1}} & \mathbf{0}_{\left(\tilde{t}-q_{1}-q_{2}\right) \times p_{2}} & \mathbf{0}_{\left(\tilde{t}-q_{1}-q_{2}\right) \times p_{b}}
\end{array}\right] \cdot R_{R}\left(\boldsymbol{\theta}_{R, \beta}\right),
$$

where $R_{R}\left(\boldsymbol{\theta}_{R, \beta}\right) \in \mathbb{R}^{r \times r}$ as in Lemma 1 and $R_{H}\left(\boldsymbol{\theta}_{H}\right):=R_{H}\left(\left(\boldsymbol{\theta}_{H_{1}}, \boldsymbol{\theta}_{H_{2}}, \boldsymbol{\theta}_{H_{b}}\right)\right):=$ $R_{H_{1}}\left(\boldsymbol{\theta}_{H_{1}}\right) R_{H_{2}}\left(\boldsymbol{\theta}_{H_{2}}\right) R_{H_{b}}\left(\boldsymbol{\theta}_{H_{b}}\right) \in \mathbb{R}^{\tilde{t} \times \tilde{t}}$ is a product of Givens rotations corresponding to the entries in the blocks highlighted by bold font. The three matrices are defined as follows:

$$
\begin{aligned}
& R_{H_{1}}\left(\boldsymbol{\theta}_{H_{1}}\right):=\prod_{i=1}^{p_{1}} \prod_{j=1}^{\tilde{q}-q_{2}-r_{1}} R_{t, i, \delta_{H_{1}}(j)+j}\left(\theta_{H_{1},\left(\tilde{t}-q_{2}-r_{1}\right)(i-1)+j}\right), \delta_{H_{1}}(j):= \begin{cases}p_{1} & \text { if } j \leq q_{1}-r_{1} \\
\tilde{t}_{1}+\tilde{t}_{2}+p_{b} & \text { else, }\end{cases} \\
& R_{H_{2}}\left(\boldsymbol{\theta}_{H_{2}}\right):=\prod_{i=1}^{p_{2}} \prod_{j=1}^{\tilde{f}-q_{1}-r_{2}} R_{t, p_{1}+i, \delta_{H_{2}}(j)+j}\left(\theta_{H_{2},\left(\tilde{t}-q_{1}-r_{2}\right)(i-1)+j}\right), \delta_{H_{2}}(j):= \begin{cases}\tilde{t}_{1}+p_{2} & \text { if } j \leq q_{2}-r_{2} \\
\tilde{t}_{1}+\tilde{t}_{2}+p_{b} & \text { else, }\end{cases} \\
& R_{H_{b}}\left(\boldsymbol{\theta}_{H_{b}}\right):=\prod_{i=1}^{p_{b}} \prod_{j=1}^{\tilde{T}-q_{1}-q_{2}} R_{t, p_{1}+p_{2}+i, \tilde{1}_{1}+\tilde{t}_{2}+p_{b}+j}\left(\theta_{\left.H_{b},\left(\tilde{t}-q_{1}-q_{2}\right)(i-1)+j\right) .}\right.
\end{aligned}
$$

Consequently, a parameterization of the orthocomplement of the cointegrating space is based on the mapping:

$$
\begin{aligned}
& \mathcal{C}_{1}^{r}\left(\boldsymbol{\theta}_{H}, \boldsymbol{\theta}_{R, \mathcal{C}}\right):= \\
& {\left[\tilde{H} \cdot R_{H}\left(\boldsymbol{\theta}_{H}\right)^{\prime}\left[\begin{array}{ccc}
0_{p_{1} \times\left(q_{1}-r_{1}\right)} & 0_{p_{1} \times\left(q_{2}-r_{2}\right)} & 0_{p_{1} \times\left(\tilde{t}-q_{1}-q_{2}\right)} \\
I_{q_{1}-r_{1}} & 0_{\left(q_{1}-r_{1}\right) \times\left(q_{2}-r_{2}\right)} & 0_{\left(q_{1}-r_{1}\right) \times\left(\tilde{t}-q_{1}-q_{2}\right)} \\
0_{p_{2} \times\left(q_{1}-r_{1}\right)} & 0_{p_{2} \times\left(q_{2}-r_{2}\right)} & 0_{p_{2} \times\left(\tilde{t}-q_{1}-q_{2}\right)} \\
0_{\left(q_{2}-r_{2}\right) \times\left(q_{1}-r_{1}\right)} & I_{q_{2}-r_{2}} & 0_{\left(q_{2}-r_{2}\right) \times\left(\tilde{t}-q_{1}-q_{2}\right)} \\
0_{p_{b} \times\left(q_{1}-r_{1}\right)} & 0_{p_{b} \times\left(q_{2}-r_{2}\right)} & 0_{p_{b} \times\left(\tilde{t}-q_{1}-q_{2}\right)} \\
0_{\left(\tilde{t}-q_{1}-q_{2}\right) \times\left(q_{1}-r_{1}\right)} & 0_{\left(\tilde{t}-q_{1}-q_{2}\right) \times\left(q_{2}-r_{2}\right)} & I_{\tilde{t}-q_{1}-q_{2}}
\end{array}\right], \tilde{H}_{\perp}\right] \cdot R_{R}\left(\boldsymbol{\theta}_{R, \mathcal{C})}\right),}
\end{aligned}
$$

where $R_{H}\left(\boldsymbol{\theta}_{H}\right) \in \mathbb{R}^{\tilde{\tau} \times \tilde{t}}$ as above and $R_{R}\left(\boldsymbol{\theta}_{R, \mathcal{C}}\right) \in \mathbb{R}^{(s-r) \times(s-r)}$ as in Lemma 1. Please note that for all $\boldsymbol{\theta}_{H}$, $\boldsymbol{\theta}_{R, \beta}$ and $\boldsymbol{\theta}_{R, \mathcal{C}}$ it holds that $\beta^{r}\left(\boldsymbol{\theta}_{H}, \boldsymbol{\theta}_{R, \beta}\right)^{\prime} \mathcal{C}_{1}^{r}\left(\boldsymbol{\theta}_{H}, \boldsymbol{\theta}_{R, \mathcal{C}}\right)=0_{r \times(s-r)}$. The number of parameters restricted under $H_{0}^{\prime \prime}$ is equal to $r_{1}\left(q_{1}-r_{1}\right)+r_{2}\left(q_{2}-r_{2}\right)+\left(r_{1}+r_{2}\right)\left(\tilde{t}-q_{1}-q_{2}\right)+(s-r)(s-r+1) / 2$, and thus, through $q_{1}$ and $q_{2}$, depends on the dimension $t_{b}$ of the intersection of the columns spaces of $H_{1}$ and $H_{2}$. The reduction of the number of free parameters matches the degrees of freedom of the test statistics in Johansen (1995, Theorem 7.5), if $\beta$ is identified, which is the case if $r_{1} \leq \tilde{t}_{1}$ and $r_{2} \leq \tilde{t}_{2}$.

Using the mapping $\beta^{r}(\cdot)$ as a basis for a parameterization allows to introduce another type of hypotheses of the form:
(iv) $H_{0}^{\prime \prime \prime}: \beta_{\perp}=\mathcal{C}_{1}=\left[H_{1} \varphi_{1}, \ldots, H_{c} \varphi_{c}\right], \beta_{\perp} \in \mathbb{R}^{s \times(s-r)}, H_{j} \in O_{s, t_{j}}, \varphi_{j} \in O_{t_{j}, r_{j}}, r_{j} \leq t_{j} \leq s$, for $j=1, \ldots, c$ such that $\sum_{j=1}^{c} r_{j}=s-r$. The ortho-complement of the cointegrating space is contained in the column spaces of the (full rank) matrices $H_{k}$.

This type of hypothesis allows, e.g., to test for the presence of cross-unit cointegrating relations (cf. Wagner and Hlouskova 2009, Definition 1) in, e.g., multi-country data sets.

Hypotheses on the cointegrating space at frequency $\omega_{k}=\pi$ can be treated analogously to hypotheses on the cointegrating space at frequency $\omega_{k}=0$.

Testing hypotheses on cointegrating spaces at frequencies $0<\omega_{k}<\pi$ has to be discussed in more detail, as one also has to consider the space spanned by PCIVs, compare Example 2. There are $2\left(s-d_{1}^{k}\right)$ linearly independent PCIVs of the form $\beta(z)=\beta_{0}+\beta_{1} z$. Every PCIV corresponds to a vector $z_{k} \beta_{0}+\beta_{1} \in \mathbb{C}^{s}$ orthogonal to $\mathcal{C}_{k}$ and consequently hypotheses on the space spanned by PCIVs can be transformed to hypotheses on the complex column space of $\mathcal{C}_{k} \in \mathbb{C}^{s \times d_{1}^{k}}$.

Consider, e.g., an extension of the first type of hypothesis of the form

$$
\begin{aligned}
H_{0}^{k}:\left[\begin{array}{ll}
\gamma_{0} & \tilde{\gamma}_{0} \\
\gamma_{1} & \tilde{\gamma}_{1}
\end{array}\right] & =\left[\begin{array}{cc}
I_{s} & 0_{s \times s} \\
-\cos \left(\omega_{k}\right) I_{s} & \sin \left(\omega_{k}\right) I_{s}
\end{array}\right]\left[\begin{array}{cc}
\left(\tilde{H}_{0} \tilde{\phi}_{0}-\tilde{H}_{1} \tilde{\phi}_{1}\right) & \left(\tilde{H}_{0} \tilde{\phi}_{1}+\tilde{H}_{1} \tilde{\phi}_{0}\right) \\
-\left(\tilde{H}_{0} \tilde{\phi}_{1}+\tilde{H}_{1} \tilde{\phi}_{0}\right) & \left(\tilde{H}_{0} \tilde{\phi}_{0}-\tilde{H}_{1} \tilde{\phi}_{1}\right)
\end{array}\right] \\
& =\left[\begin{array}{cc}
I_{s} & 0_{s \times s} \\
-\cos \left(\omega_{k}\right) I_{s} & \sin \left(\omega_{k}\right) I_{s}
\end{array}\right]\left[\begin{array}{cc}
\tilde{H}_{0} & \tilde{H}_{1} \\
-\tilde{H}_{1} & \tilde{H}_{0}
\end{array}\right]\left[\begin{array}{cc}
\tilde{\phi}_{0} & \tilde{\phi}_{1} \\
-\tilde{\phi}_{1} & \tilde{\phi}_{0}
\end{array}\right],
\end{aligned}
$$

with $\tilde{H}_{0}, \tilde{H}_{1} \in \mathbb{R}^{s \times t}, \tilde{\phi}_{0}, \tilde{\phi}_{1} \in \mathbb{R}^{t \times r}, r \leq t<s$, which implies that the column space of $\mathcal{C}_{k}$ is equal to the orthocomplement of the column space of $\left(\tilde{H}_{0}+i \tilde{H}_{1}\right)\left(\tilde{\phi}_{0}+i \tilde{\phi}_{1}\right)$. This general hypothesis encompasses, e.g., the hypothesis $\left[\gamma_{0}^{\prime}, \gamma_{1}^{\prime}\right]^{\prime}=H \phi=\left[H_{0}^{\prime}, H_{1}^{\prime}\right]^{\prime} \phi$, with $H \in \mathbb{R}^{2 s \times t}, H_{0}, H_{1} \in \mathbb{R}^{s \times t}, \phi \in \mathbb{R}^{t \times r}$, by setting $\tilde{\phi}_{0}:=\tilde{\phi}_{1}:=\tilde{\phi}, \tilde{H}_{0}:=H_{0}$ and $\tilde{H}_{1}:=-\left(\cos \left(\omega_{k}\right) H_{0}+H_{1}\right) / \sin \left(\omega_{k}\right)$. The extension is tailored to include the pairwise structure of PCIVs and to simplify transformation into hypotheses on the complex matrix $\mathcal{C}_{k}$ used in the parameterization. The parameterization of the set of matrices corresponding to $H_{0}^{k}$ is derived from a mapping of the form given in (15), with $\check{R}_{L}\left(\check{\boldsymbol{\theta}}_{L}\right)$ and $R_{R}\left(\theta_{R}\right)$ replaced by $\breve{Q}_{L}\left(\check{\boldsymbol{\varphi}}_{L}\right):=\prod_{i=1}^{t-r} \prod_{j=1}^{r} Q_{t, i, t-r+j}\left(\varphi_{L, r(i-1)+j}\right) \in \mathbb{R}^{t \times t}$ and $D_{d}\left(\boldsymbol{\varphi}_{D}\right) Q_{R}\left(\boldsymbol{\varphi}_{R}\right)$ as in Lemma 2.

Similarly, the three other types of hypotheses on the cointegrating spaces considered above can be extended to hypotheses on the space of PCIVs in the MFI(1) case. They translate into hypotheses on complex valued matrices $\beta_{k}$ orthogonal to $\mathcal{C}_{k}$. To parameterize the set of matrices restricted according to these null hypotheses, Lemma 2 is used. Thus, the restrictions implied by the extensions of all four types of hypotheses to hypotheses on the dynamic cointegrating spaces at frequencies $0<\omega_{k}<\pi$ for $\mathrm{MFI}(1)$ processes can be implemented using Givens rotations.

A different case of interest is the hypothesis of at least $m$ linearly independent CIVs $b_{j} \in \mathbb{R}^{s}$, $j=1, \ldots, m$ with $0<m \leq s-d_{1}^{k}$, i.e., an $m$-dimensional static cointegrating space at frequency $0<\omega_{k}<\pi$, which we discuss as another illustrative example to the procedure for the case of cointegration at complex unit roots.

For the dynamic cointegrating space, this hypothesis implies the existence of $2 m$ linearly independent PCIVs of the form $\beta_{1}(z)=b_{j}$ and $\beta_{2}(z)=b_{j} z, j=1, \ldots, m$. In light of the discussion above the necessary condition for these two polynomials to be PCIVs is equivalent to $b_{j}^{\prime} \mathcal{C}_{k}=0$, for $j=1, \ldots, m$. This restriction is similar to $H_{0}^{\prime}$ discussed above, except for the fact that the cointegrating vectors $b_{j}$ are not fully specified. This hypothesis is equivalent to the existence of an $m$-dimensional real kernel of $\mathcal{C}_{k}$. A suitable parameterization is derived from the following mapping

$$
C\left(\boldsymbol{\theta}_{b}, \boldsymbol{\varphi}\right):=R_{L}\left(\boldsymbol{\theta}_{b}\right)\left[\begin{array}{c}
0_{m \times d_{1}^{k}} \\
C_{U}(\boldsymbol{\varphi})
\end{array}\right],
$$

where $\boldsymbol{\theta}_{b} \in[0,2 \pi)^{m(s-m)}$ and $C_{U}(\boldsymbol{\varphi}):=C_{U}\left(\boldsymbol{\varphi}_{L}, \boldsymbol{\varphi}_{D}, \boldsymbol{\varphi}_{R}\right) \in U_{s-m, d_{1}^{k}}$ as in Lemma 2. The difference in the number of free parameters without restrictions and with restrictions is equal to $m(s-m)$.

The hypotheses can also be tested jointly for the cointegrating spaces of several unit roots.

### 5.1.3. Testing Hypotheses on the Adjustment Coefficients

As in the case of hypotheses on the cointegrating spaces $\beta_{k}$, hypotheses on the adjustment coefficients $\alpha_{k}$ are typically formulated as hypotheses on the column spaces of $\alpha_{k}$. We only focus on hypotheses on the real valued $\alpha_{1}$ corresponding to frequency zero. Analogous hypotheses may be considered for $\alpha_{k}$ at frequencies $\omega_{k} \neq 0$, using the same ideas.

The first type of hypothesis on $\alpha_{1}$ is of the form $H_{\alpha}: \alpha_{1}=A \psi, A \in \mathbb{R}^{s \times t}, \psi \in \mathbb{R}^{t \times r}$ and therefore, can be rewritten as $\mathcal{B}_{1} A \psi=0$. W.l.o.g. let $A \in O_{s, t}$ and $A_{\perp} \in O_{s, s-t}$. We deal with this type of hypothesis as with $H_{0}: \beta=H \varphi$ in the previous section by simply reversing the roles of $\mathcal{C}_{1}$ and $\mathcal{B}_{1}$. We, therefore, consider the set of feasible matrices $\mathcal{B}_{1}^{\prime}$ as a subset in $O_{s, s-r}$ and use the mapping $\mathcal{B}_{1}^{\prime}\left(\check{\boldsymbol{\theta}}_{L}, \boldsymbol{\theta}_{R}\right)=\left[A \check{R}_{L}\left(\check{\boldsymbol{\theta}}_{L}\right)^{\prime}\left[I_{t-r}, 0_{r \times(t-r)}\right]^{\prime}, A_{\perp}\right] R_{R}\left(\boldsymbol{\theta}_{R}\right)$ to derive a parameterization, while $\mathcal{C}_{1}^{\prime}$ is restricted to be a p.u.t. matrix and the set of feasible matrices $\mathcal{C}_{1}^{\prime}$ is parameterized accordingly.

As a second type of hypothesis Juselius (2006, sct. 11.9, p. 200) discusses $H_{\alpha}^{\prime}: \alpha_{1, \perp}=H \psi$, $H \in \mathbb{R}^{s \times t}, \psi \in \mathbb{R}^{t \times(s-r)}$, linked to the absence of permanent effects of shocks $H_{\perp} \varepsilon_{t}$ on any of the variables of the system. Assume w.l.o.g. $H_{\perp} \in O_{s, s-t}$. Using the parameterization of $O_{s-r}\left(H_{\perp}\right)$ defined in (13) for the set of feasible matrices $\mathcal{B}_{1}^{\prime}$ and the parameterization of the set of p.u.t. matrices for the set of feasible matrices $\mathcal{C}_{1}^{\prime}$, implements this restriction.

The restrictions on $H_{\alpha}$ reduce the number of free parameters by $r(s-t)$ and the restrictions implied by $H_{\alpha}^{\prime}$ lead to a reduction by $t(s-r)$ free parameters, compared to the unrestricted case,
which matches in both cases the number of degrees of freedom of the corresponding test statistic in the VECM framework.

### 5.1.4. Restrictions on the Deterministic Components

Including an unrestricted constant in the VECM equation $\Delta_{0} y_{t}=\varepsilon_{t}+\Phi_{0}$ leads to a linear trend in the solution process $y_{t}=\sum_{j=1}^{t}\left(\varepsilon_{j}+\Phi_{0}\right)+y_{1}=\sum_{j=1}^{t} \varepsilon_{j}+y_{1}+\Phi_{0} t$, for $t>1$. If one restricts the constant to $\Phi_{0}=\alpha \tilde{\Phi}_{0}, \tilde{\Phi}_{0} \in \mathbb{R}^{r}$ in a general VECM equation as given in (4), with $\Pi=\alpha \beta^{\prime}$ of rank $r$, no summation to linear trends in the solution process occurs, while a constant non-zero mean is still present in the cointegrating relations, i.e., the process $\left\{\beta^{\prime} y_{t}\right\}_{t \in \mathbb{Z}}$. Analogously an unrestricted linear trend $\Phi_{1} t$ in the VECM equation leads to a quadratic trend of the form $\Phi_{1} t(t-1) / 2$ in the solution process, which is excluded by the restriction $\Phi_{1} t=\alpha \tilde{\Phi}_{1} t$.

In the VECM framework, compare Johansen (1995, sct. 5.7, p. 81), five restrictions related to the coefficients corresponding to the constant and the linear trend are commonly considered:

1. $H(r): \Phi d_{t}=\Phi_{1} t+\Phi_{0}$, i.e., unrestricted constant and linear trend,
2. $H^{*}(r): \Phi d_{t}=\alpha \tilde{\Phi}_{1} t+\Phi_{0}$, i.e., unrestricted constant, linear trend restricted to cointegrating relations,
3. $H_{1}(r): \Phi d_{t}=\Phi_{0}, \quad$ i.e., unrestricted constant, no linear trend,
4. $H_{1}^{*}(r): \Phi d_{t}=\alpha \tilde{\Phi}_{0}, \quad$ i.e., constant restricted to cointegrating relations, no linear trend,
5. $H_{2}(r): \Phi d_{t}=0, \quad$ i.e., no deterministic components present,
with $\Phi_{0}, \Phi_{1} \in \mathbb{R}^{s}$ and $\tilde{\Phi}_{0}, \tilde{\Phi}_{1}, \in \mathbb{R}^{r}$ and the following consequences for the solution processes: Under $H(r)$ the solution process contains a quadratic trend in the direction of the common trends, i.e., in $\left\{\beta_{\perp}^{\prime} y_{t}\right\}_{t \in \mathbb{Z}}$, and a linear trend in the direction of the cointegrating relations, i.e., in $\left\{\beta^{\prime} y_{t}\right\}_{t \in \mathbb{Z}}$. Under $H^{*}(r)$ the quadratic trend is not present. $H_{1}(r)$ features a linear trend only in the directions of the common trends, $H_{2}(r)$ a constant only in these directions. Under $H_{1}^{*}(r)$ the constant is also present in the directions of the cointegrating relations.

In the state space framework the deterministic components can be added in the output equation $y_{t}=\mathcal{C} x_{t}+\Phi d_{t}+\varepsilon_{t}$, compare (9). Consequently, the above considered hypotheses can be imposed by formulating linear restrictions on $\Phi$. These can be directly parameterized by including the following deterministic components in the five considered cases:

1. $H(r): \Phi d_{t}=\mathcal{C}_{1} \tilde{\Phi}_{2} t^{2}+\Phi_{1} t+\Phi_{0}$,
2. $H^{*}(r): \Phi d_{t}=\Phi_{1} t+\Phi_{0}$,
3. $H_{1}(r): \Phi d_{t}=\mathcal{C}_{1} \tilde{\Phi}_{1} t+\Phi_{0}$,
4. $H_{1}^{*}(r): \Phi d_{t}=\Phi_{0}$,
5. $H_{2}(r): \Phi d_{t}=\mathcal{C}_{1} \tilde{\Phi}_{0}$,
where $\Phi_{0}, \Phi_{1} \in \mathbb{R}^{s}$ and $\tilde{\Phi}_{0}, \tilde{\Phi}_{1}, \tilde{\Phi}_{2} \in \mathbb{R}^{d_{1}^{1}}$. The component $\mathcal{C}_{1} \tilde{\Phi}_{0}$ captures the influence of the initial value $\mathcal{C}_{1} x_{1,1}$ in the output equation.

In the VECM framework for the seasonal MFI(1) case, with $\Pi_{k}=\alpha_{k} \beta_{k}^{\prime}$ of rank $r_{k}$ for $0<\omega_{k}<\pi$, the deterministic component usually includes restricted seasonal dummies of the form $\alpha_{k} \tilde{\Phi}_{k} z_{k}^{t}+$ $\overline{\alpha_{k}} \tilde{\Phi}_{k}\left(z_{k}\right)^{t}, \tilde{\Phi}_{k} \in \mathbb{C}^{r_{k}}$ to avoid summation in the directions of the stochastic trends. The state space framework allows to straightforwardly include seasonal dummies in the output equation in the form of $\Phi_{k} z_{k}^{t}+\overline{\Phi_{k}\left(z_{k}\right)^{t}}, \Phi_{k} \in \mathbb{C}^{s}$. Again, it is of interest whether these components are unrestricted or whether they take the form of $\mathcal{C}_{k} \tilde{\Phi}_{k} z_{k}^{t}+\overline{\mathcal{C}_{k}} \tilde{\Phi}_{k}\left(z_{k}\right)^{t}, \tilde{\Phi}_{k} \in \mathbb{C}^{d_{1}^{k}}$, similarly allowing for a reinterpretation of these components as influence of the initial values $x_{1, k}$ on the output.

Please note that $\Phi_{k} z_{k}^{t}+\overline{\Phi_{k}\left(z_{k}\right)^{t}}$ is equivalently given by $\breve{\Phi}_{k, 1} \sin \left(\omega_{k} t\right)+\breve{\Phi}_{k, 2} \cos \left(\omega_{k} t\right)$ using real coefficients $\breve{\Phi}_{k, 1}, \breve{\Phi}_{k, 2} \in \mathbb{R}^{s}$ and the desired restrictions can be implemented accordingly.

### 5.2. The $I(2)$ Case

The state space unit root structure of $\mathrm{I}(2)$ processes is of the form $\Omega_{S}=\left(\left(0, d_{1}^{1}, d_{2}^{1}\right)\right)$, where the integer $d_{1}^{1}$ equals the dimension of $x_{t, 1}^{E}$, and $d_{2}^{1}$ equals the dimension of $\left[\left(x_{t, 2}^{G}\right)^{\prime},\left(x_{t, 2}^{E}\right)^{\prime}\right]^{\prime}$. Recall that the solution for $t>0$ and $x_{1, u}=0$ of the system in canonical form in this setting is given by

$$
\begin{aligned}
y_{t}= & \mathcal{C}_{1,1}^{E} x_{t, 1}^{E}+\mathcal{C}_{1,2}^{G} x_{t, 2}^{G}+\mathcal{C}_{1,2}^{E} x_{t, 2}^{E}+\mathcal{C}_{\bullet} x_{t, \bullet}+\Phi d_{t}+\varepsilon_{t} \\
= & \mathcal{C}_{1,1}^{E} \mathcal{B}_{1,2,1} \sum_{k=1}^{t-1} \sum_{j=1}^{k} \varepsilon_{t-j}+\left(\mathcal{C}_{1,1}^{E} \mathcal{B}_{1,1}+\mathcal{C}_{1,2}^{G} \mathcal{B}_{1,2,1}+\mathcal{C}_{1,2}^{E} \mathcal{B}_{1,2,2}\right) \sum_{j=1}^{t-1} \varepsilon_{t-j} \\
& \quad+\mathcal{C}_{\bullet} \sum_{j=1}^{t-1} \mathcal{A}_{\bullet}^{j-1} \mathcal{B} \bullet \varepsilon_{t-j}+\mathcal{C}_{\bullet} \mathcal{A}_{\bullet}^{t-1} x_{1, \bullet}+\Phi d_{t}+\varepsilon_{t}
\end{aligned}
$$

For VAR processes integrated of order two the integers $d_{1}^{1}$ and $d_{2}^{1}$ of the corresponding state space unit root structure are linked to the ranks of the matrices $\Pi=\alpha \beta^{\prime}$ (denoted as $r=r_{0}$ ) and $\alpha_{\perp}^{\prime} \Gamma \beta_{\perp}=\xi \eta^{\prime}$ (denoted as $m=r_{1}$ ) in the VECM setting, as discussed in Section 2. It holds that $r=s-d_{2}^{1}$ and $m=d_{2}^{1}-d_{1}^{1}$. The relation of the state space unit root structure to the cointegration indices $r_{0}, r_{1}, r_{2}$ was also discussed in Section 3.

Again, both the integers $d_{1}^{1}$ and $d_{2}^{1}$ and the ranks $r$ and $m$, and consequently also the indices $r_{0}, r_{1}$ and $r_{2}$, are closely related to the dimensions of the spaces spanned by CIVs and PCIVs. In the $I(2)$ case the static cointegrating space of order $((0,2),(0,1))$ is the orthocomplement of the column space of $\mathcal{C}_{1,1}^{E}$ and thus of dimension $s-d_{1}^{1}$. The dimension of the space spanned by CIVs of order $((0,2),\{ \})$ is equal to $s-d_{2}^{1}-r_{c, G}$, where $r_{c, G}$ denotes the rank of $\mathcal{C}_{1,2}^{G}$, since this space is the orthocomplement of the column space of $\left[\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{G}, \mathcal{C}_{1,2}^{E}\right]$. The space spanned by the PCIVs $\beta_{0}+\beta_{1} z$ of order $((0,2),\{ \})$ is of dimension smaller or equal to $2 s-d_{1}^{1}-d_{2}^{1}$, due to the orthogonality constraint on $\left[\beta_{0}^{\prime}, \beta_{1}^{\prime}\right]^{\prime}$ given in Example 3.

Consider the matrices $\beta, \beta_{1}$ and $\beta_{2}$ as defined in Section 2. From a state space realization $(\mathcal{A}, \mathcal{B}, \mathcal{C})$ in canonical form corresponding to a VAR process it immediately follows that the columns of $\beta_{2}$ span the same space as the columns of the sub-block $\mathcal{C}_{1,1}^{E}$. The same relation holds true for $\beta_{1}$ and the sub-block $\mathcal{C}_{1,2}^{E}$. With respect to polynomial cointegration, Bauer and Wagner (2012) show that the rank of $\mathcal{C}_{1,2}^{G}$ determines the number of minimum degree polynomial cointegrating relations, as discussed in Example 3. If $\mathcal{C}_{1,2}^{G}=0$, then there exists no vector $\gamma$, such that $\left\{\gamma^{\prime} y_{t}\right\}_{t \in \mathbb{Z}}$ is integrated and cointegrated with $\left\{\beta_{2}^{\prime} \Delta_{0} y_{t}\right\}_{t \in \mathbb{Z}}$. In this case $\left\{\beta^{\prime} y_{t}\right\}_{t \in \mathbb{Z}}$ is a stationary process.

The deterministic components included in the $\mathrm{I}(2)$ setting are typically a constant and a linear trend. As in the $\operatorname{MFI}(1)$ case, identifiability problems occur, if we consider a non-zero initial state $x_{1, u}$ : The solution to the state space equations for $t>0$ and $x_{1, u} \neq 0$ is given by:

$$
y_{t}=\sum_{j=1}^{t-1} \mathcal{C} \mathcal{A}^{j-1} \mathcal{B} \varepsilon_{t-j}+\mathcal{C}_{1,1}^{E}\left(x_{1,1}^{E}+x_{1,2}^{G}(t-1)\right)+\mathcal{C}_{1,2}^{G} x_{1,2}^{G}+\mathcal{C}_{1,2}^{E} x_{1,2}^{E}+\mathcal{C}_{\bullet} \mathcal{A}_{\bullet}^{t-1} x_{1, \bullet}+\Phi d_{t}+\varepsilon_{t}
$$

Hence, if $\Phi d_{t}=\Phi_{0}+\Phi_{1} t$, the output equation contains the terms $\mathcal{C}_{1,1}^{E} x_{1,1}^{E}+\mathcal{C}_{1,2}^{G} x_{1,2}^{G}+\mathcal{C}_{1,2}^{E} x_{1,2}^{E}-$ $\mathcal{C}_{1,1}^{E} x_{1,2}^{G}+\Phi_{0}$ and $\left(\mathcal{C}_{1,1}^{E} x_{1,2}^{G}+\Phi_{1}\right)$ t. Again, this implies non-identifiability, which is resolved by assuming $x_{1, u}=0$, compare Remark 6 .

### 5.2.1. Testing Hypotheses on the State Space Unit Root Structure

To simplify notation we use

$$
\bar{M}\left(d_{1}^{1}, d_{2}^{1}\right):= \begin{cases}\overline{M\left(\left(\left(0, d_{1}^{1}, d_{2}^{1}\right)\right), n-d_{1}^{1}-d_{2}^{1}\right)} & \text { if } d_{1}^{1}>0 \\ \overline{M\left(\left(\left(0, d_{2}^{1}\right)\right), n-d_{2}^{1}\right)} & \text { if } d_{1}^{1}=0, d_{2}^{1}>0 \\ \overline{M_{\bullet}, n} & \text { if } d_{1}^{1}=d_{2}^{1}=0\end{cases}
$$

with $n \geq d_{1}^{1}+d_{2}^{1}$. Here $\bar{M}\left(d_{1}^{1}, d_{2}^{1}\right)$ for $d_{1}^{1}+d_{2}^{1}>0$ denotes the closure of the set of transfer functions of order $n$ that possess a state space unit root structure of either $\Omega_{S}=\left(\left(0, d_{1}^{1}, d_{2}^{1}\right)\right)$ or $\Omega_{S}=\left(\left(0, d_{2}^{1}\right)\right)$ in case of $d_{1}^{1}=0$, while $\bar{M}(0,0)$ denotes the closure of the set of all stable transfer functions of order $n$.

Considering the relations between the different sets of transfer functions given in Corollary 4 shows that the following relations hold (assuming $s \geq 4$; the columns are arranged to include transfer functions with the same dimension of $\mathcal{A}_{u}$ ):


Please note that $\bar{M}\left(d_{1}^{1}, d_{2}^{1}\right)$ corresponds to $H_{s-d_{2}^{1}, d_{2}^{1}-d_{1}^{1}}=H_{r, r_{1}}$ in Johansen (1995). Therefore, the relationships between the subsets match the ones in Johansen (1995, Table 9.1) and the ones found by Jensen (2013). The latter type of inclusions appear for instance for $\bar{M}(0,2)$, containing transfer functions corresponding to $I(1)$ processes, which is a subset of the set $\bar{M}(1,0)$ of transfer functions corresponding to $I(2)$ processes.

The same remarks as in the $\mathrm{MFI}(1)$ case also apply in the $\mathrm{I}(2)$ case: When testing for $H_{0}: \Omega_{S}=$ $\left(\left(0, d_{1,0}^{1}, d_{2,0}^{1}\right)\right)$, all attainable state space unit root structures $\mathcal{A}\left(\left(\left(0, d_{1,0}^{1}, d_{2,0}^{1}\right)\right)\right)$ have to be included in the null hypothesis.

### 5.2.2. Testing Hypotheses on CIVs and PCIVs

Johansen (2006) discusses several types of hypotheses on the cointegrating spaces of different orders. These deal with properties of $\beta$, joint properties of $\left[\beta, \beta_{1}\right]$ or the occurrence of non-trivial polynomial cointegrating relations. Boswijk and Paruolo (2017), moreover, discuss testing hypotheses on the loading matrices of common trends (corresponding in our setting to testing hypotheses on $\mathcal{C}_{1}$ ).

We commence with hypotheses of the form $H_{0}: \beta=K \varphi$ and $H_{0}^{\prime}: \beta=[b, \varphi]$ just as in the MFI(1) case at unit root one, since hypotheses on $\beta$ correspond to hypotheses on its orthocomplement spanned by $\left[\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{E}\right]$ in the VARMA framework:

Hypotheses of the form $H_{0}: \beta=K \varphi, K \in \mathbb{R}^{s \times t}, \varphi \in \mathbb{R}^{t \times r}$ imply $\varphi^{\prime} K^{\prime}\left[\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{E}\right]=0$. W.l.o.g. let $K \in O_{s, t}$ and $K_{\perp} \in O_{s, s-t}$. As in the parameterization under $H_{0}$ in the $\operatorname{MFI}(1)$ case at unit root one, compare (15), use the mapping

$$
\left[\mathcal{C}_{1,1}^{E, r}, \mathcal{C}_{1,2}^{E, r}\right]\left(\check{\boldsymbol{\theta}}_{L}, \boldsymbol{\theta}_{R}\right):=\left[K \cdot \check{R}_{L}\left(\check{\boldsymbol{\theta}}_{L}\right)^{\prime}\left[\begin{array}{c}
I_{t-r} \\
0_{r \times(t-r)}
\end{array}\right], \quad K_{\perp}\right] \cdot R_{R}\left(\boldsymbol{\theta}_{R}\right)
$$

to derive a parameterization of the set of feasible matrices $\left[\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{E}\right]$, i.e., a joint parameterization of both sets of matrices $\mathcal{C}_{1,1}^{E}$ and $\mathcal{C}_{1,2}^{E}$, where $\left[\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{E}\right] \in O_{s, s-r}$.

Hypotheses of the form $H_{0}^{\prime}: \beta=[b, \varphi], b \in \mathbb{R}^{s \times t}, \varphi \in \mathbb{R}^{s \times(r-t)}, 0<t \leq r$ are equivalent to $b^{\prime}\left[\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{E}\right]=0$. Assume w.l.o.g. $b \in O_{s, t}$ and parameterize the set of feasible matrices $\mathcal{C}_{1,1}^{E}$ using $O_{s, d_{1}^{1}}(b)$ as defined in (13) and the set of feasible matrices $\mathcal{C}_{1,2}^{E}$ using $O_{s, d_{2}^{1}-d_{1}^{1}}\left(\left[b, C_{1,1}^{E}\right]\right)$. Alternatively, parameterize the set of feasible matrices jointly as elements $\left[\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{E}\right] \in O_{s, s-r}(b)$.

Applications using the VECM framework allow for testing hypotheses on $\left[\beta, \beta_{1}\right]$. In the VARMA framework, these correspond to hypotheses on the orthogonal complement of $\left[\beta, \beta_{1}\right]$, i.e., $\mathcal{C}_{1,1}^{E}$. Implementation of different types of hypotheses on $\left[\beta, \beta_{1}\right]$ proceeds as for similar hypotheses on $\beta$ in the $\operatorname{MFI}(1)$ case at unit root one, replacing $\left[\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{E}\right]$ by $\mathcal{C}_{1,1}^{E}$.

The hypothesis of no minimum degree polynomial cointegrating relations implies the restriction $\mathcal{C}_{1,2}^{G}=0$, compare Example 3. Therefore, we can test all hypotheses considered in Johansen (2006) also in our more general setting.

### 5.2.3. Testing Hypotheses on the Adjustment Coefficients

Hypotheses on $\alpha$ and $\xi$ as defined in (6) and (7) correspond to hypotheses on the spaces spanned by the rows of $\mathcal{B}_{1,2,1}$ and $\mathcal{B}_{1,2,2}$. For VAR processes integrated of order two, the row space of $\mathcal{B}_{1,2,1}$ is equal to the orthogonal complement of the column space of $[\alpha, \alpha \perp \xi]$, while the row space of $\mathcal{B}_{1,2}:=\left[\mathcal{B}_{1,2,1}^{\prime}, \mathcal{B}_{1,2,2}^{\prime}\right]^{\prime}$ is equal to the orthogonal complement of the column space of $\alpha$. The restrictions corresponding to hypotheses on $\alpha$ and $\xi$ can be implemented analogously to the restrictions corresponding to hypotheses on $\alpha_{1}$ in Section 5.1.3, reversing the roles of the relevant sub-blocks in $B_{u}$ and $\mathcal{C}_{u}$ accordingly.

### 5.2.4 Restrictions on the Deterministic Components

The $\mathrm{I}(2)$ case is, with respect to the modeling of deterministic components, less well studied than the $\operatorname{MFI}(1)$ case. In most theory papers they are simply left out, with the notable exception Rahbek et al. (1999), dealing with the inclusion of a constant term in the I(2)-VECM representation. The main reason for this appears to be the way deterministic components in the defining vector error correction representation translate into deterministic components in the corresponding solution process. An unrestricted constant in the VECM for $I(2)$ processes leads to a linear trend in $\left\{\beta_{1}^{\prime} y_{t}\right\}_{t \in \mathbb{Z}}$ and a quadratic trend in $\left\{\beta_{2}^{\prime} y_{t}\right\}_{t \in \mathbb{Z}}$, while an unrestricted linear trend results in quadratic and cubic trends in the respective directions. Already in the $I(1)$ case discussed above five different cases-with respect to integration and asymptotic behavior of estimators and tests-need to be considered separately. An all encompassing discussion of the restrictions on the coefficients of a constant and a linear trend in the $I(2)$ case requires the specification of even more cases. As an alternative approach in the VECM framework, deterministic components could be dealt with by replacing $y_{t}$ with $y_{t}-\Phi d_{t}$ in the VECM equation. This has recently been considered in Johansen and Nielsen (2018) and is analogous to our approach in the state space framework.

As before, in the $\operatorname{MFI}(1)$ or $\mathrm{I}(1)$ case, the analysis of (the impact of) deterministic components is straightforward in the state space framework, which effectively stems from their additive inclusion in the Granger-type representation, compare (9). Choose, e.g., $\Phi d_{t}=\Phi_{0}+\Phi_{1} t$, as in the $\mathrm{I}(1)$ case. In analogy to Section 5.1.4, linear restrictions of deterministic components in relation to the static and polynomial cointegrating spaces can be embedded in a parameterization. Focusing on $\Phi_{0}$, e.g., this is achieved by

$$
\Phi_{0}=\left[\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{E}\right] \phi_{0}+\tilde{\mathcal{C}}_{1,2} \tilde{\phi}_{0}+\mathcal{C}_{\perp} \check{\phi}_{0}
$$

where the columns of $\tilde{\mathcal{C}_{1,2}}$ are a basis for the column space of $\mathcal{C}_{1,2}^{G}$, which does not necessarily have full column rank, and the columns of $\mathcal{C}_{\perp}$ span the orthocomplement of the column space of $\left[\mathcal{C}_{1,1}^{E}, \mathcal{C}_{1,2}^{E}, \tilde{\mathcal{C}_{1,2}}\right]$. The matrix $\Phi_{1}$ can be decomposed analogously. The corresponding parametrization then allows to consider different restricted versions of deterministic components and to study the asymptotic behavior of estimators and tests for these cases.

## 6. Summary and Conclusions

Vector autoregressive moving average (VARMA) processes, which can be cast equivalently in the state space framework, may be useful for empirical analysis compared to the more restrictive class of vector autoregressive (VAR) processes for a variety of reasons. These include invariance with respect to marginalization and aggregation, parsimony as well as the fact that the log-linearized solutions to DSGE models are typically VARMA processes rather than VAR processes. To realize the potential of these advantages necessitates, in our view, to develop cointegration analysis for VARMA processes to a
similar extent as it is developed for VAR processes. The necessary first steps of this research agenda are to develop a set of structure theoretical results that allow subsequently developing statistical inference procedures. Bauer and Wagner (2012) provides the very first step of this agenda by providing a canonical form for unit root processes in the state space framework, which is shown in that paper to be very convenient for cointegration analysis.

Based on the earlier canonical form paper this paper derives a state space model parameterization for VARMA processes with unit roots using the state space framework. The canonical form and a fortiori the parameterization based on it are constructed to facilitate the investigation of the unit root and (static and polynomial) cointegration properties of the considered process. Furthermore, the paper shows that the framework allows to test a large variety of hypotheses on cointegrating ranks and spaces, clearly a key aspect for the usefulness of any method to analyze cointegration. In addition to providing general results, throughout the paper all results are discussed in detail for the multiple frequency $I(1)$ and $I(2)$ cases, which cover the vast majority of applications.

Given the fact that (as shown in Hazewinkel and Kalman 1976) VARMA unit root processes cannot be continuously parameterized, the set of all unit root processes (as defined in this paper) is partitioned according to a multi-index $\Gamma$ that includes the state space unit root structure. The parameterization is shown to be a diffeomorphism on the interior of the considered sets. The topological relationships between the sets forming the partitioning of all transfer functions considered are studied in great detail for three reasons: First, pseudo maximum likelihood estimation effectively amounts to maximizing the pseudo likelihood function over the closures of sets of transfer functions, $\bar{M}_{\Gamma}$ in our notation. Second, related to the first item, the relations between subsets of $M_{\Gamma}$ have to be understood in detail as knowledge concerning these relations is required for developing (sequential) pseudo likelihood-ratio tests for the numbers of stochastic trends or cycles. Third, of particular importance for the implementation of, e.g., pseudo maximum likelihood estimators, we discuss the existence of generic pieces.

In this respect we derive two results: First, for correctly specified state space unit root structure and system order of the stable subsystem -and thus correctly specified system order-we explicitly describe generic indices $\Gamma_{g}\left(\Omega_{S}, n_{\bullet}\right)$ such that $M_{\Gamma_{g}\left(\Omega_{S}, n_{\bullet}\right)}$ is open and dense in the set of all transfer functions with state space unit root structure $\Omega_{S}$ and system order of the stable subsystem $n_{\bullet}$. This result forms the basis for establishing consistent estimators of the transfer functions-and via continuity of the parameterization-of the parameter estimators when the state space unit root structure and system order are known. Second, in case only an upper bound on the system order is known (or specified), we show the existence of a generic multi-index $\Gamma_{\alpha_{0,8}(n)}$ for which the set of corresponding transfer functions $M_{\Gamma_{\alpha \bullet g(n)}}$ is open and dense in the set $\bar{M}_{n}$ of all non-explosive transfer functions whose order (or McMillan degree) is bounded by $n$. This result is the basis for consistent estimation (on an open and dense subset) when only an upper bound of the system order is known. In turn this estimator is the starting point for determining $\Omega_{S}$, using the subset relationships alluded to above in the second point. For the $\operatorname{MFI}(1)$ and $\mathrm{I}(2)$ cases we show in detail that similar subset relations (concerning cointegrating ranks) as in the cointegrated VAR MFI(1) and I(2) cases hold, which suggests constructing similar sequential test procedures for determining the cointegrating ranks as in the VAR cointegration literature.

Section 5 is devoted to a detailed discussion of testing hypotheses on the cointegrating spaces, again for both the $\mathrm{MFI}(1)$ and the $\mathrm{I}(2)$ case. In this section, particular emphasis is put on modeling deterministic components. The discussion details how all usually formulated and tested hypotheses concerning (static and polynomial) cointegrating vectors, potentially in combination with (un-)restricted deterministic components, in the VAR framework can also be investigated in the state space framework.

Altogether, the paper sets the stage to develop pseudo maximum likelihood estimators, investigate their asymptotic properties (consistency and limiting distributions) and tests based on them for determining cointegrating ranks that allow performing cointegration analysis for cointegrated

VARMA processes. The detailed discussion of the MFI(1) and I(2) cases benefits the development of statistical theory dealing with these cases undertaken in a series of companion papers.

Author Contributions: The authors of the paper have contributed equally, via joint efforts, regarding both ideas, research, and writing. Conceptualization, all authors; methodology, all authors; formal analysis, P.d.M.R. and L.M.; investigation, all authors; writing-original draft preparation, P.d.M.R. and L.M.; writing-review and editing, all authors.; project administration, D.B. and M.W.; funding acquisition, D.B. and M.W. All authors have read and agreed to the published version of the manuscript.
Funding: This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation-Projektnummer 276051388) which is gratefully acknowledged. We acknowledge support for the publication costs by the Deutsche Forschungsgemeinschaft and the Open Access Publication Fund of Bielefeld University.
Acknowledgments: We thank the editors, Rocco Mosconi and Paolo Paruolo, as well as anonymous referees for helpful suggestions. The views expressed in this paper are solely those of the authors and not necessarily those of the Bank of Slovenia or the European System of Central Banks. On top of this the usual disclaimer applies.
Conflicts of Interest: The authors declare no conflict of interest.

## Appendix A. Proofs of the Results of Section 3

Appendix A.1. Proof of Lemma 1
(i) Let $C_{j}$ be a sequence in $O_{s, d}$ converging to $C_{0}$ for $j \rightarrow \infty$. By continuity of matrix multiplication

$$
C_{0}^{\prime} C_{0}=\left(\lim _{j \rightarrow \infty} C_{j}\right)^{\prime} \lim _{j \rightarrow \infty} C_{j}=\lim _{j \rightarrow \infty}\left(C_{j}^{\prime} C_{j}\right)=I_{d}
$$

Thus, $C_{0} \in O_{s, d}$, which shows that $O_{s, d}$ is closed. By construction $\left[C^{\prime} C\right]_{i, i}=\sum_{j=1}^{s} c_{j, i}^{2}$ Since $\left[C^{\prime} C\right]_{i, i}=1$ for all $C \in O_{s, d}$ and $i=1, \ldots, d$, the entries of $C$ are bounded.
(ii) By definition $C_{O}(\boldsymbol{\theta})$ is a product of matrices whose elements are either constant or infinitely often differentiable functions of the elements of $\boldsymbol{\theta}$.
(iii) The algorithm discussed above Lemma 1 maps every $C \in O_{s, d}$ to $\left[I_{d}, 0_{s-d \times d}^{\prime}\right]^{\prime}$. Since $R_{q, i, j}(\theta)^{-1}=$ $R_{q, i, j}(\theta)^{\prime}$ for all $q, i, j$ and $\theta, C$ can be obtained by multiplying $\left[I_{d}, 0_{s-d \times d}^{\prime}\right]^{\prime}$ with the transposed Givens rotations.
(iv) As discussed, $C_{O}^{-1}(\cdot)$ is obtained from a repeated application of the algorithm described in Remark 10. In each step two entries are transformed to polar coordinates. According to Amann and Escher (2008, chp. 8, p. 204) the transformation to polar coordinates is infinitely often differentiable with infinitely often differentiable inverse for $\theta>0$ (and hence $r>0$ ), i.e., on the interior of the interval $[0, \pi)$. Thus, $C_{O}^{-1}$ is a concatenation of functions which are infinitely often differentiable on the interior of $\Theta_{O}^{\mathbb{R}}$ and is thus infinitely often differentiable, if $\theta_{j}>0$ for all components of $\theta$.
Clearly, the interior of $\Theta_{O}^{\mathbb{R}}$ is open and dense in $\Theta_{O}^{\mathbb{R}}$. By the definition of continuity the pre-image of the interior of $\Theta_{O}^{\mathbb{R}}$ is open in $O_{s, d}$. By (iii) there exists a $\theta_{0}$ for arbitrary $C_{0} \in O_{s, d}$ such that $C_{O}\left(\boldsymbol{\theta}_{0}\right)=C_{0}$. Since the interior of $\Theta_{O}^{\mathbb{R}}$ is dense in $\Theta_{O}^{\mathbb{R}}$ there exists a sequence $\theta_{j}$ in the interior of $\Theta_{O}^{\mathbb{R}}$ such that $\boldsymbol{\theta}_{j} \rightarrow \boldsymbol{\theta}_{0}$. Then $C_{O}\left(\boldsymbol{\theta}_{j}\right) \rightarrow C_{0}$ because of the continuity of $C_{O}$. Since $C_{O}\left(\boldsymbol{\theta}_{j}\right)$ is a sequence in the pre-image of the interior of $\Theta_{O}^{\mathbb{R}}$, it follows that the pre-image of the interior of $\Theta_{O}^{\mathbb{R}}$ is dense in $O_{s, d}$.
(v) For any $C \in O_{s, s}$ it holds that $1=\operatorname{det}\left(C^{\prime} C\right)=\operatorname{det}(C)^{2}$ and $\operatorname{det}(C) \in \mathbb{R}$, which implies $\operatorname{det}(C) \in$ $\{-1,1\}$. Since the determinant is a continuous function on square matrices, both sets $O_{s, s}^{+}$and $O_{s, S}^{-}$are disjoint and closed.
(vi) The proof proceeds analogously to the proof of (iii).
(vii) A function defined on two disjoint subsets is infinitely often differentiable if and only if the two functions restricted to the subsets are infinitely often differentiable. The same arguments as used in (iv) together with the results in (ii) imply that $C_{O}^{-1}: O_{s, s}^{+} \rightarrow \Theta_{O}^{\mathbb{R}}$ and $\left.C_{O}^{ \pm}(\cdot)\right|_{S_{s, s}^{+}}$are infinitely often differentiable with infinitely often differentiable inverse on an open subset of
$O_{s, s}^{+}$. Clearly, the multiplication with $I_{s}^{-}$is infinitely often differentiable with infinitely often differentiable inverse, which implies that $\left.C_{O}^{ \pm}(\cdot)\right|_{O_{s, s}^{-}}$is infinitely often differentiable with infinitely often differentiable inverse on an open subset of $O_{s, s}^{-}$, from which the result follows.

## Appendix A.2. Proof of Lemma 2

(i) Let $C_{j}$ be a sequence in $U_{s, d}$ converging to $C_{0}$ for $j \rightarrow \infty$. By continuity of matrix multiplication

$$
C_{0}^{\prime} C_{0}=\left(\lim _{j \rightarrow \infty} C_{j}\right)^{\prime} \lim _{j \rightarrow \infty} C_{j}=\lim _{j \rightarrow \infty}\left(C_{j}^{\prime} C_{j}\right)=I_{d}
$$

Thus, $C_{0} \in U_{s, d}$, which shows that $U_{s, d}$ is closed. By construction $\left[C^{\prime} C\right]_{i, i}=\sum_{j=1}^{s}\left|c_{j, i}\right|^{2}$. Since $\left[C^{\prime} C\right]_{i, i}=1$ for all $C \in U_{s, d}$ and $i=1, \ldots, d$, the entries of $C$ are bounded.
(ii) By definition $C_{U}(\boldsymbol{\varphi})$ is a product of matrices whose elements are either constant or infinitely often differentiable functions of the elements of $\varphi$.
(iii) The algorithm discussed above Lemma 2 maps every $C \in U_{s, d}$ to $\left[D_{d}\left(\boldsymbol{\varphi}_{D}\right), 0_{s-d \times d}^{\prime}\right]^{\prime}$ with $D_{d}\left(\boldsymbol{\varphi}_{D}\right)=\operatorname{diag}\left(e^{i \varphi_{D, 1}}, \ldots, e^{i \varphi_{D, d}}\right)$. Since $Q_{q, i, j}(\varphi)^{-1}=Q_{q, i, j}(\varphi)^{\prime}$ for all $q, i, j$ and $\varphi, C$ can be obtained by multiplying $\left[D_{d}\left(\boldsymbol{\varphi}_{D}\right), 0_{s-d \times d}^{\prime}\right]^{\prime}$ with the transposed Givens rotations.
(iv) The algorithms in Remark 12 and above Lemma 2 describe $C_{U}^{-1}$ in detail. The determination of an element of $\boldsymbol{\varphi}_{L}$ or $\boldsymbol{\varphi}_{R}$ uses the transformation of two complex numbers into polar coordinates in step 2 of Remark 12, which according to Amann and Escher (2008, chp. 8, p. 204) is infinitely often differentiable with infinitely often differentiable inverse except for non-negative reals, which are the complement of an open and dense subset of the complex plane. Step 3 of Remark 12 uses the formulas $\varphi_{1}=\tan ^{-1}\left(\frac{b}{a}\right)$, which is infinitely often differentiable for $a>0$, and $\varphi_{2}=$ $\varphi_{a}-\varphi_{b} \bmod 2 \pi$, which is infinitely often differentiable for $\varphi_{a} \neq \varphi_{b}$, which occurs on an open and dense subset of $[0,2 \pi) \times[0,2 \pi)$. For the determination of an element of $\varphi_{D}$ a complex number of modulus one is transformed in polar coordinates which is infinitely often differentiable on an open and dense subset of complex numbers of modulus one compare again Amann and Escher (2008, chp. 8, p. 204). Thus, $C_{U}^{-1}$ is a concatenation of functions which are infinitely often differentiable on open and dense subsets of their domain of definition and is thus infinitely often differentiable on an open and dense subset of $U_{s, d}$.

## Appendix A.3. Proof of Theorem 2

(i) The multi-index $\Gamma$ is unique for a transfer function $k \in M_{n}$, since it only contains information encoded in the canonical form. Therefore, $M_{\Gamma}$ is well defined. Since conversely for every transfer function $k \in M_{n}$ a multi-index $\Gamma$ can be found, $M_{\Gamma}$ constitutes a partitioning of $M_{n}$. Furthermore, using the canonical form, it is straightforward to see that the mapping attaching the triple $(\mathcal{A}, \mathcal{B}, \mathcal{C}) \in \Delta_{\Gamma}$ in canonical form to a transfer function $k \in M_{\Gamma}$ is homeomorphic (bijective, continuous, with continuous inverse): Bijectivity is a consequence of the definition of the canonical form. $T_{p t}$ continuity of the transfer function as a function of the matrix triples is obvious from the definition of $T_{p t}$. Continuity of the inverse can be shown by constructing the canonical form starting with an overlapping echelon form (which is continuous according to Hannan and Deistler 1988, chp. 2) and subsequently transforming the state basis to reach the canonical form. This involves the calculation of a Jordan normal form with fixed structure. This is an analytic mapping (cf. Chatelin 1993, Theorem 4.4.3). Finally, the restrictions on $C$ and $B$ are imposed. For given multi-index $\Gamma$ these transformations are continuous (as discussed above they involve QR decompositions to obtain unitary block columns for the blocks of $C$, rotations to p.u.t form with fixed structure for the blocks of $B$ and transformations to echelon canonical form for the stable part).
(ii) The construction of the triple $(\mathcal{A}(\theta), \mathcal{B}(\theta), \mathcal{C}(\theta))$ for given $\theta$ and $\Gamma$ is straightforward: $\mathcal{A}_{u}$ is uniquely determined by $\Gamma$. Since $\boldsymbol{\theta}_{B, p}$ contains the entries of $\mathcal{B}_{u}$ restricted to be positive and $\boldsymbol{\theta}_{B, f}$
contains the free parameters of $\mathcal{B}_{u}$, the mapping $\boldsymbol{\theta}_{B, p} \times \boldsymbol{\theta}_{B, f} \rightarrow \mathcal{B}_{u}$ is continuous. The mapping $\theta_{\bullet} \rightarrow\left(\mathcal{A}_{\bullet}, \mathcal{B}_{\bullet}, \mathcal{C}_{\bullet}\right)$ is continuous (cf. Hannan and Deistler 1988, Theorem 2.5.3 (ii)). The mapping $\boldsymbol{\theta}_{C, E} \times \boldsymbol{\theta}_{C, G} \rightarrow \mathcal{C}_{u}$ consists of iterated applications of $C_{O}$, and $C_{U}$ (compare Lemmas 1 and 2) which are differentiable and thus continuous and iterated applications of the extensions of the mappings $C_{O, d_{2}-d_{1}}$ and $C_{O, G}$ (compare Corollaries 1 and 2 ) to general unit root structures and to complex matrices. The proof that these functions are differentiable is analogous to the proofs of Lemma 1 and Lemma 2.
(iii) The definitions of $\boldsymbol{\theta}_{B, f}$ and $\boldsymbol{\theta}_{B, p}$ immediately imply that they depend continuously on $\mathcal{B}_{u}$. The parameter vector $\theta_{\bullet}$ depends continuously on $\left(\mathcal{A}_{\bullet}, \mathcal{B}_{\bullet}, \mathcal{C}_{\bullet}\right)$ (cf. Hannan and Deistler 1988, Theorem 2.5.3 (ii)). The existence of an open and dense subset of matrices $\mathcal{C}_{u}$ such that the mapping attaching parameters to the matrices is continuous follows from arguments contained in the proofs of Lemmas 1 and 2.

## Appendix B. Proofs of the Results of Section 4

## Appendix B.1. Proof of Theorem 3

For the first inclusion the proof can be divided into two parts, discussing the stable and the unstable subsystem separately. The result with regard to the stable subsystem is due to Hannan and Deistler (1988, Theorem 2.5.3 (iv)). For the unstable subsystem $\left(\tilde{\Omega}_{S}, \tilde{p}\right) \leq\left(\Omega_{S}, p\right)$ implies the existence of a matrix $S$ as described in Definition 9. Partition $S=\left[\begin{array}{l}S_{1} \\ S_{2}\end{array}\right]$ such that $S_{1} p=p_{1} \geq \tilde{p}$. Let $\tilde{k}$ be an arbitrary transfer function in $M_{\tilde{\Gamma}}=\pi\left(\Delta_{\tilde{\Gamma}}\right)$ with corresponding state space realization $(\tilde{\mathcal{A}}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}}) \in \Delta_{\tilde{\Gamma}}$. Then, we find matrices $B_{1}$ and $C_{1}$ such that for the state space realization given by $\mathcal{A}=S\left[\begin{array}{cc}\tilde{\mathcal{A}} & \tilde{j}_{12} \\ 0 & \tilde{J}_{2}\end{array}\right] S^{\prime}, \mathcal{B}=S\left[\begin{array}{c}\tilde{\mathcal{B}} \\ B_{1}\end{array}\right]$ and $\mathcal{C}=\left[\begin{array}{cc}\tilde{\mathcal{C}} & C_{1}\end{array}\right] S^{\prime}$ it holds that $(\mathcal{A}, \mathcal{B}, \mathcal{C}) \in \Delta_{\Gamma}$. Then, $\left(\mathcal{A}_{j}, \mathcal{B}_{j}, \mathcal{C}_{j}\right)=\left(\mathcal{A}, S \operatorname{diag}\left(I_{n_{1}}, j^{-1} I_{n_{2}}\right) S^{\prime} \mathcal{B}, \mathcal{C}\right) \in \Delta_{\Gamma}$, where $n_{i}$ is the number of rows of $S_{i}$ for $i=1,2$ converges for $j \rightarrow \infty$ to $\left(\mathcal{A}, S\left[\begin{array}{c}\tilde{\mathcal{B}} \\ 0\end{array}\right], \mathcal{C}\right) \in \overline{\Delta_{\Gamma}}$, which is observationally equivalent to $(\tilde{\mathcal{A}}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}})$. Consequently, $\tilde{k}=\pi\left(\mathcal{A}, S\left[\begin{array}{c}\tilde{\mathcal{B}} \\ 0\end{array}\right], \mathcal{C}\right) \in \pi\left(\overline{\Delta_{\Gamma}}\right)$.

To show the second inclusion, consider a sequence of systems $\left(\mathcal{A}_{j}, \mathcal{B}_{j}, \mathcal{C}_{j}\right) \in \Delta_{\Gamma}, j \in \mathbb{N}$ converging to $\left(A_{0}, B_{0}, C_{0}\right) \in \overline{\Delta_{\Gamma}}$. We need to show $\bar{\Gamma} \in \bigcup_{\tilde{\Gamma} \in \mathcal{K}(\Gamma)}\{\check{\Gamma} \leq \tilde{\Gamma}\}$, where $\bar{\Gamma}$ is the multi-index corresponding to $\left(A_{0}, B_{0}, C_{0}\right)$.

For the stable system we can separate the subsystem $\left(A_{j, s}, B_{j, s}, C_{j, s}\right)$ remaining stable in the limit and the part with eigenvalues of $A_{j}$ tending to the unit circle. As discussed in Section 4.1.2, ( $A_{j, s}, B_{j, s}, C_{j, s}$ ) converges to the stable subsystem ( $A_{0, \bullet}, B_{0, \bullet}, C_{0, \bullet}$ ) whose Kronecker indices can only be smaller than or equal to $\alpha_{\bullet}$ (cf. Hannan and Deistler 1988, Theorem 2.5.3).

The remaining subsystem consists of the unstable subsystem of $\left(\mathcal{A}_{j}, \mathcal{B}_{j}, \mathcal{C}_{j}\right)$ which converges to $\left(A_{0, u}, B_{0, u}, C_{0, u}\right)$ and the second part of the stable subsystem containing all stable eigenvalues of $A_{j}$ converging to the unit circle. The limiting combined subsystem $\left(A_{0, c}, B_{0, c}, C_{0, c}\right)$ is such that $A_{0, c}$ is block diagonal. If the limiting combined subsystem is minimal and $B_{0, u}$ has a structure corresponding to $p$, this shows that the pair $\left(\bar{\Omega}_{S}, \bar{p}\right)$ extends $\left(\Omega_{S}, p\right)$ in accordance with the definition of $\mathcal{K}(\Gamma)$.

Since the limiting subsystem is not necessarily minimal and $B_{0, u}$ has not necessarily a structure corresponding to $p$, eliminating coordinates of the state and adapting the corresponding structure indices $p$ may result in a pair $\left(\bar{\Omega}_{S}, \bar{p}\right)$ that is smaller than the pair $\left(\tilde{\Omega}_{S}, \tilde{p}\right)$ corresponding to an element of $\mathcal{K}(\Gamma)$.

## Appendix B.2. Proof of Theorem 4

The multi-index $\Gamma$ contains three components: $\Omega_{S}, p_{,} \alpha_{\bullet}$. For given $\Omega_{S}$ the selection of the structures indices $p_{\max }$ introducing the fewest restrictions, such that in its boundary all possible p.u.t. matrices occur, was discussed in Section 4.2. Choosing this maximal element $p_{\max }$ then implies that
all systems of given state space unit root structure correspond to a multi-index that is smaller than or equal to $\left(\Omega_{S}, p_{\max }, \beta_{\bullet}\right)$, where $\beta_{\bullet}$ is a Kronecker index corresponding to state space dimension $n_{\bullet}$. For the Kronecker indices of order $n_{\bullet}$ it is known that there exists one index $\alpha_{\bullet, g}$ such that $M_{\alpha_{\bullet, g}}$ is open and dense in $\overline{M_{n_{\bullet}}}$. The set $M_{\Omega_{S}, p_{\max }, \beta_{\bullet}}$ is, therefore, contained in $\overline{M_{\Omega_{S}, p_{\max }, \alpha_{\bullet}, g}}$ which implies (14) with $\Gamma_{g}\left(\Omega_{S}, n_{\bullet}\right):=\left(\Omega_{S}, p_{\max }, \alpha_{\bullet}, g\right)$.

For the second claim choose an arbitrary state space realization $(\mathcal{A}, \mathcal{B}, \mathcal{C})$ in canonical form such that $\pi(\mathcal{A}, \mathcal{B}, \mathcal{C}) \in M\left(\Omega_{S}, n_{\bullet}\right)$ for arbitrary $\Omega_{S}$. Define the sequence $\left(A_{j}, B_{j}, C_{j}\right)_{j \in \mathbb{N}}$ by $A_{j}=\left(1-j^{-1}\right) \mathcal{A}$, $B_{j}=\left(1-j^{-1}\right) \mathcal{B}, C_{j}=\mathcal{C}$. Then $\lambda_{|\max |}\left(A_{j}\right)<1$ holds for all $j$, which implies $\pi\left(A_{j}, B_{j}, C_{j}\right) \in \overline{M_{\Gamma_{0 \bullet g}(n)}}$ for every $n \geq n_{u}\left(\Omega_{s}\right)+n_{\bullet}$ and every $j$. The continuity of $\pi$ implies $\pi(\mathcal{A}, \mathcal{B}, \mathcal{C})=\lim _{j \rightarrow \infty} \pi\left(A_{j}, B_{j}, C_{j}\right) \in$ $\overline{M_{\Gamma_{0, g}(n)}}$.

## Appendix B.3. Proof of Theorem 5

(i) Assume that there exists a sequence $k_{i} \in \overline{M_{\Gamma}}$ converging to a transfer function $k_{0} \in M_{\Gamma}$. For such a sequence the size of the Jordan blocks for every unit root are identical from some $i_{0}$ onwards since eigenvalues depend continuously on the matrices (cf. Chatelin 1993): Thus, the stable part of the transfer functions $k_{i}$ must converge to the stable part of the transfer function $k_{0}$, since the sum of the algebraic multiplicity of all eigenvalues inside the open unit disc cannot drop in the limit. Since $V_{\alpha}$ (the set of all stable transfer functions with Kronecker index $\alpha$ ) is open in $\overline{V_{\alpha}}$ according to Hannan and Deistler (1988, Theorem 2.5.3) this implies that the stable part of $k_{i}$ has Kronecker index $\alpha_{\bullet}$ from some $i_{0}$ onwards.

For the unstable part of the transfer function note that in $M_{\Gamma}$ for every unit root $z_{j}$ the rank of $\left(A-z_{j} I_{n}\right)^{r}$ is equal for every $r$. Thus, the maximum over $\overline{M_{\Gamma}}$ cannot be larger due to lower semi-continuity of the rank. It follows that for $k_{i} \rightarrow k_{0}$ the ranks of $\left(A-z_{j} I_{n}\right)^{r}$ for all $\left|z_{j}\right|=1$ and for all $r \in \mathbb{N}_{0}$ are identical to the ranks corresponding to $k_{0}$ from some point onwards showing that $k_{i}$ has the same state space unit root structure as $k_{0}$ from some $i_{0}$ onwards. Finally, the p.u.t. structure of sub-blocks of $\mathcal{B}_{k}$ clearly introduces an open set being defined via strict inequalities. This shows that $k_{i} \in M_{\Gamma}$ from some $i_{0}$ onwards implying that $M_{\Gamma}$ is open in $\overline{M_{\Gamma}}$.
(ii) The first inclusion was shown in Theorem 3. Comparing Definitions 10 and 11 we see $\bigcup_{\tilde{\Gamma} \in \mathcal{K}\left(\Gamma_{g}\right)} M_{\tilde{\Gamma}} \subset \bigcup_{\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right) \in \mathcal{A}\left(\Omega_{S}, n_{\bullet}\right)} M\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right)$. By the definition of the partial ordering (compare Definition 9) $\bigcup_{\tilde{\Gamma} \leq \Gamma_{g}} M_{\tilde{\Gamma}} \subset \bigcup_{\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right) \leq\left(\Omega_{S}, n_{\bullet}\right)} M\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right)$ holds. Together these two statements imply the second inclusion.
$\bigcup_{\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right) \in \mathcal{A}\left(\Omega_{S}, n_{\bullet}\right)} \bigcup_{\left(\check{\Omega}_{S}, \check{n}_{\bullet}\right) \leq\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right)} M\left(\check{\Omega}_{S}, \check{n}_{\bullet}\right) \subset \bar{M}_{\Gamma_{g}\left(\Omega_{s}, n_{\bullet}\right)}$ is a consequence of the following two statements:
(a) If $M\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right) \subset \overline{M\left(\Omega_{S}, n_{\bullet}\right)}$, then $\bigcup_{\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right) \leq\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right)} M\left(\check{\Omega}_{S}, \check{n}_{\bullet}\right) \subset \overline{M\left(\Omega_{S}, n_{\bullet}\right)}$.
(b) If $\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right) \in \mathcal{A}\left(\Omega_{S}, n_{\bullet}\right)$, then $M\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right) \subset \overline{M\left(\Omega_{S}, n_{\bullet}\right)}$.

For (a) note that for an arbitrary transfer function $\check{k} \in M\left(\check{\Omega}_{S}, \check{n}_{\bullet}\right)$ with $\left(\check{\Omega}_{S}, \check{n}_{\bullet}\right) \leq\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right)$ there is a multi-index $\check{\Gamma}$ such that $\check{k} \in M_{\check{\Gamma}}$. By the definition of the partial ordering (compare Definition 9) we find a multi-index $\tilde{\Gamma} \geq \check{\Gamma}$ such that $M_{\tilde{\Gamma}} \subset M\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right)$. By Theorem 3 and the continuity of $\pi$ we have $M_{\check{\Gamma}} \subset \pi\left(\overline{\Delta_{\tilde{\Gamma}}}\right) \subset \overline{M_{\tilde{\Gamma}}}$. Since $\overline{M\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right)} \subset \overline{M\left(\Omega_{S}, n_{\bullet}\right)}$ by assumption, $\check{k} \in \overline{M_{\tilde{\Gamma}}} \subset \overline{M\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right)} \subset \overline{M\left(\Omega_{S}, n_{\bullet}\right)}$ which finishes the proof of (a).
With respect to (b) note that by Definition $11, \mathcal{A}\left(\Omega_{S}, n_{\bullet}\right)$ contains transfer functions with two types of state space unit root structures. First, $\mathcal{A}_{u}$ corresponding to state space unit root $\tilde{\Omega}_{S}$ may be of the form

$$
S \tilde{\mathcal{A}}_{u} S^{\prime}=\left[\begin{array}{cc}
\mathcal{A}_{u} & J_{12}  \tag{A1}\\
0 & J_{2}
\end{array}\right]
$$

Second, $\check{\mathcal{A}}_{u}$ corresponding to state space unit $\operatorname{root} \check{\Omega}_{S}$ may be of the form (A1) where off-diagonal elements of $\mathcal{A}_{u}$ are replaced by zero. To prove (b) we need to show that for both cases the corresponding transfer function is contained in $\overline{M\left(\Omega_{S}, n_{\bullet}\right)}$.

We start by showing that in the second case the transfer function $\check{k}$ is contained in $\overline{M\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right)}$, where $\tilde{\Omega}_{S}$ is the state space unit root structure corresponding to $\tilde{\mathcal{A}}_{u}$ in (A1). For this, consider the sequence

$$
A_{j}=\left[\begin{array}{cc}
1 & j^{-1} \\
0 & 1
\end{array}\right], \quad B_{j}=\left[\begin{array}{l}
B_{1} \\
B_{2}
\end{array}\right], C_{j}=\left[\begin{array}{ll}
C_{1} & C_{2}
\end{array}\right] .
$$

Clearly, every system $\left(A_{j}, B_{j}, C_{j}\right)$ corresponds to an $I(2)$ process, while the limit for $j \rightarrow \infty$ corresponds to an $I(1)$ process. This shows that it is possible in the limit to trade one $I(2)$ component with two $I(1)$ components leading to more transfer functions in the $T_{p t}$ closure of $M_{\Gamma_{g}\left(\Omega_{S}, n_{\bullet}\right)}$ than only the ones included in $\pi\left(\overline{\left.\Delta_{\Gamma_{g}\left(\Omega_{S}, n_{\bullet}\right)}\right)}\right.$, where the off-diagonal entry in $A_{j}$ is restricted to equal one and hence the corresponding sequence of systems in the canonical form diverges to infinity. In a sense these systems correspond to "points at infinity": For the example given above we obtain the canonical form

$$
\mathcal{A}_{j}=\left[\begin{array}{ll}
1 & 1 \\
0 & 1
\end{array}\right], \quad \mathcal{B}_{j}=\left[\begin{array}{c}
B_{1} \\
B_{2} / j
\end{array}\right], \mathcal{C}_{j}=\left[\begin{array}{ll}
C_{1} & j C_{2}
\end{array}\right] .
$$

Thus, the corresponding parameter vector for the entries in $\mathcal{B}_{j, 2}$ converges to zero and the ones corresponding to $\mathcal{C}_{j, 2}$ to infinity.

Generalizing this argument shows that every transfer function corresponding to a pair $\left(\check{\Omega}_{S}, \check{n}_{\bullet}\right)$ in $\mathcal{A}\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right)$, where $\mathcal{A}_{u}$ can be obtained by replacing off-diagonal entries of $\mathcal{A}_{u}$ with zero, can be reached from within $M\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right)$.
To prove $\tilde{k} \in \overline{M\left(\Omega_{S}, n_{\bullet}\right)}$ in the first case, where the state space unit root structure is extended as visible in Equation (A1), consider the sequence:

$$
\tilde{A}_{j}=\left[\begin{array}{cc}
1 & 1 \\
0 & 1-j^{-1}
\end{array}\right], \quad \tilde{B}_{j}=\left[\begin{array}{c}
B_{1} \\
B_{2}
\end{array}\right], \quad \tilde{C}_{j}=\left[\begin{array}{ll}
C_{1} & C_{2}
\end{array}\right]
$$

corresponding to the following system in canonical form (except that the stable subsystem is not necessarily in echelon canonical form)

$$
\tilde{\mathcal{A}}_{j}=\left[\begin{array}{cc}
1 & 0 \\
0 & 1-j^{-1}
\end{array}\right], \quad \tilde{\mathcal{B}}_{j}=\left[\begin{array}{c}
B_{1}+j B_{2} \\
-j B_{2}
\end{array}\right], \quad \tilde{\mathcal{C}}_{j}=\left[\begin{array}{ll}
C_{1} & C_{1}-C_{2} / j
\end{array}\right] .
$$

This sequence shows that there exists a sequence of transfer functions corresponding to $I(1)$ processes with one common trend that converge to a transfer function corresponding to an $I(2)$ system. Again, in the canonical form this cannot happen as there the $(1,2)$ entry of $\tilde{A}_{j}$ would be restricted to be equal to zero. At the same time note that the dimension of the stable system is reduced due to one component of the state changing from the stable to the unit root part.

Now for a unit root structure $\tilde{\Omega}_{S}$ such that $\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right) \in \mathcal{A}\left(\Omega_{S}, n_{\bullet}\right)$, satisfying

$$
S \tilde{\mathcal{A}}_{u} S^{\prime}=\left[\begin{array}{cc}
\mathcal{A}_{u} & J_{12} \\
0 & J_{2}
\end{array}\right]
$$

the Jordan blocks corresponding to $\Omega_{S}$ are sub-blocks of the ones corresponding to $\tilde{\Omega}_{S}$, potentially involving a reordering of coordinates using the permutation matrix $S$. Taking as the approximating sequence of transfer functions $\tilde{k}_{j} \in M_{\Gamma_{g}\left(\Omega_{s}, n_{\bullet}\right)} \rightarrow k_{0} \in M_{\Gamma_{g}\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right)}$ that have the same structure $\tilde{\Omega}_{S}$ but replacing $J_{2}$ by $\frac{j-1}{j} J_{2}$ leads to processes with state space unit root structure $\Omega_{S}$.

For the stable part of $\tilde{k}_{j}$ we can separate the part containing poles tending to the unit circle (contained in $J_{2}$ ) and the remaining transfer function $\tilde{k}_{j, s}$, which has Kronecker indices $\tilde{\alpha} \leq \alpha_{\bullet}$. However, the results of Hannan and Deistler (1988, Theorem 2.5.3) then imply that the limit remains in $\overline{M_{\alpha_{\bullet}}}$ and hence allows for an approximating sequence in $M_{\alpha_{\bullet}}$.

Both results combined constitute the whole set of attainable state space unit root structures in Definition 11 and prove (b).
As follows from Corollary $4, \overline{M\left(\Omega_{S}, n_{\bullet}\right)}=\overline{M_{\Gamma_{g}\left(\Omega_{S}, n_{\bullet}\right)}}$. Thus, (b) implies $\bigcup_{\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right) \in \mathcal{A}\left(\Omega_{S}, n_{\bullet}\right)} M\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right) \subset \overline{M_{\Gamma_{g}\left(\Omega_{S}, n_{\bullet}\right)}}$ and (a) adds the second union showing the subset inclusion.
It remains to show equality for the last set inclusion. Thus, we need to show that for $k_{j} \in M_{\Gamma_{g}\left(\Omega_{S}, n_{\bullet}\right)}, k_{j} \rightarrow k_{0}$, it holds that $k_{0} \in M\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right)$, where $\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right) \leq\left(\check{\Omega}_{S}, \check{n}_{\bullet}\right) \in \mathcal{A}\left(\Omega_{S}, n_{\bullet}\right)$. To this end note that the rank of a matrix is a lower semi-continuous function such that for a sequence of matrices $E_{j}$ with limit $E_{0}$, we have

$$
\operatorname{rank}\left(\lim _{j \rightarrow \infty} E_{j}\right)=\operatorname{rank}\left(E_{0}\right) \leq \lim \inf _{j \rightarrow \infty} \operatorname{rank}\left(E_{j}\right) .
$$

Then, consider a sequence $k_{j}(z) \in M_{\Gamma_{g}\left(\Omega_{s}, n_{\bullet}\right)}, j \in \mathbb{N}$. We can find a converging sequence of systems $\left(A_{j}, B_{j}, C_{j}\right)$ realizing $k_{j}(z)$. Therefore, choosing $E_{j}=\left(A_{j}-z_{k} I_{n}\right)^{r}$ we obtain that

$$
\operatorname{rank}\left(\left(A_{0}-z_{k} I_{n}\right)^{t}\right) \leq n-\sum_{r=1}^{t} d_{j, h_{k}-r+1}^{k}
$$

since $k_{j}(z) \in M_{\Gamma_{g}\left(\Omega_{s}, n_{\bullet}\right)}$ implies that the number $d_{j, h_{k}-r+1}^{k}$ of the generalized eigenvalues at the unit roots is governed by the entries of the state space unit root structure $\Omega_{s}$. This implies that $\sum_{r=1}^{t} d_{j, h_{k}-r+1}^{k} \leq \sum_{r=1}^{t} d_{0, h_{k}-r+1}^{k}$ for $t=1,2, \ldots, n$. Consequently, the limit has at least as many chains of generalized eigenvalues of each maximal length as dictated by the state space unit root structure $\Omega_{S}$ for each unit root of the limiting system.
Rearranging the rows and columns of the Jordan normal form using a permutation matrix $S$ it is then obvious that either the limiting matrix $A_{0}$ has additional eigenvalues, where thus

$$
S \mathcal{A}_{0} S^{\prime}=\left[\begin{array}{cc}
\mathcal{A}_{j} & \tilde{J}_{12} \\
0 & \tilde{J}_{2}
\end{array}\right]
$$

must hold. Or upper diagonal entries in $\mathcal{A}_{j}$ must be changed from ones to zeros in order to convert some of the chains to lower order. One example in this respect was given above: For $A_{j}=\left[\begin{array}{cc}1 & 1 / j \\ 0 & 1\end{array}\right]$ the rank of $\left(A_{j}-I_{2}\right)^{r}$ is equal to 1 for $r=1$ and 0 for $r=2$. For the limit we obtain $A_{0}=I_{2}$ and hence the rank is zero for $r=1,2$. The corresponding indices are $d_{j, 1}^{1}=1, d_{j, 2}^{1}=1$ for the approximating sequence and $d_{0,1}^{1}=0, d_{0,2}^{1}=2$ for the limit respectively. Summing these indices starting from the last one, one obtains $d_{j, 2}^{1}=1 \leq d_{0,2}^{1}=2$ and $d_{j, 1}^{1}+d_{j, 2}^{1}=2 \leq d_{0,1}^{1}+d_{0,2}^{1}=2$.
Hence the state space unit root structure corresponding to ( $A_{0}, B_{0}, C_{0}$ ) must be attainable according to Definition 11. The number of stable state components must decrease accordingly.

Finally, the limiting system $\left(A_{0}, B_{0}, C_{0}\right)$ is potentially not minimal. In this case the pair $\left(\tilde{\Omega}_{S}, \tilde{n}_{\bullet}\right)$ is reduced to a smaller one, concluding the proof.

## References

Amann, Herbert, and Joachim Escher. 2008. Analysis III. Basel: Birkhäuser Basel.
Aoki, Massanao. 1990. State Space Modeling of Time Series. New York: Springer.
Bauer, Dietmar, and Martin Wagner. 2003. On Polynomial Cointegration in the State Space Framework. Mimeo. [CrossRef]
Bauer, Dietmar, and Martin Wagner. 2005. Autoregressive Approximations of Multiple Frequency I(1) Processes. IHS Economics Series, Institut für Höhere Studien-Institute for Advanced Studies (IHS) Vienna, No. 174. Available online: http:/ /hdl.handle.net/10419/72306 (accessed on 3 November 2020).
Bauer, Dietmar, and Martin Wagner. 2012. A State Space Canonical Form for Unit Root Processes. Econometric Theory 28: 1313-49. [CrossRef]
Boswijk, H. Peter, and Paolo Paruolo. 2017. Likelihood Ratio Tests of Restrictions on Common Trends Loading Matrices in I(2) VAR Systems. Econometrics 5: 28. [CrossRef]
Campbell, John Y. 1994. Inspecting the Mechanism: An Analytical Approach to the Stochastic Growth Model. Journal of Monetary Economics 33: 463-506. [CrossRef]
Chatelin, Françoise. 1993. Eigenvalues of Matrices. New York: John Wiley \& Sons.
Engle, Robert F., and Clive W.J. Granger. 1987. Cointegration and Error Correction: Representation, Estimation and Testing. Econometrica 55: 251-76. [CrossRef]
Golub, Gene H., and Charles F. van Loan. 1996. Matrix Computations, 3rd ed. Baltimore: The Johns Hopkins University Press.
Granger, Clive W.J. 1981. Some Properties of Time Series Data and Their Use in Econometric Model Specification. Journal of Econometrics 16: 121-30. [CrossRef]
Hannan, Edward J., and Manfred Deistler. 1988. The Statistical Theory of Linear Systems. New York: John Wiley \& Sons.
Hazewinkel, Michiel, and Rudolf E. Kalman. 1976. Invariants, Canonical Forms and Moduli for Linear, Constant, Finite Dimensional, Dynamical Systems. In Mathematical Systems Theory. Edited by Giovanni Marchesini and Sanjoy Kumar Mitter. Berlin: Springer, chp. 4, pp. 48-60.
Jensen, Andreas N. 2013. The Nesting Structure of the Cointegrated Vector Autoregressive Models. Paper presented at the QED Conference 2013, Vienna, Austria, May 3-4 .
Johansen, Søren. 1991. Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. Econometrica 59: 1551-80. [CrossRef]
Johansen, Søren. 1995. Likelihood-Based Inference in Cointegrated Vector Auto-Regressive Models. Oxford: Oxford University Press.
Johansen, Søren. 1997. Likelihood Analysis of the I(2) Model. Scandinavian Journal of Statistics 24: 433-62. [CrossRef]
Johansen, Søren. 2006. Statistical Analysis of Hypotheses on the Cointegrating Relations in the I(2) Model. Journal of Econometrics 132: 81-115. [CrossRef]
Johansen, Søren, and Morton Ø. Nielsen. 2018. The Cointegrated Vector Autoregressive Model with General Deterministic Terms. Journal of Econometrics 202: 214-29. [CrossRef]
Johansen, Søren, and Ernst Schaumburg. 1999. Likelihood Analysis of Seasonal Cointegration. Journal of Econometrics 88: 301-39. [CrossRef]
Juselius, Katarina. 2006. The Cointegrated VAR Model: Methodology and Applications. Oxford: Oxford University Press.
Lewis, Richard, and Gregory C. Reinsel. 1985. Prediction of Multivariate Time Series by Autoregressive Model Fitting. Journal of Multivariate Analysis 16: 393-411. [CrossRef]
Otto, Markus. 2011. Rechenmethoden für Studierende der Physik im ersten Jahr. Heidelberg: Spektrum Akademischer Verlag.
Paruolo, Paolo. 1996. On the Determination of Integration Indices in I(2) Systems. Journal of Econometrics 72: 313-56. [CrossRef]
Paruolo, Paolo. 2000. Asymptotic Efficiency of the Two Stages Estimator in I(2) Systems. Econometric Theory 16: 524-50. [CrossRef]

Poskitt, Donald S. 2006. On the Identification and Estimation of Nonstationary and Cointegrated ARMAX Systems. Econometric Theory 22: 1138-75. [CrossRef]
Rahbek, Anders, Hans C. Kongsted, and Clara Jorgensen. 1999. Trend Stationarity in the I(2) Cointegration Model. Journal of Econometrics 90: 265-89. [CrossRef]
Saikkonen, Pentti. 1992. Estimation and Testing of Cointegrated Systems by an Autoregressive Approximation. Econometric Theory 8: 1-27. [CrossRef]
Saikkonen, Pentti, and Ritva Luukkonen. 1997. Testing Cointegration in Infinite Order Vector Autoregressive Processes. Journal of Econometrics 81: 93-126. [CrossRef]
Sims, Christopher A., James H. Stock, and Mark W. Watson. 1990. Inference in Linear Time Series Models with Some Unit Roots. Econometrica 58: 113-44. [CrossRef]
Wagner, Martin. 2018. Estimation and Inference for Cointegrating Regressions. In Oxford Research Encyclopedia of Economics and Finance. Oxford: Oxford University Press.
Wagner, Martin, and Jaroslava Hlouskova. 2009. The Performance of Panel Cointegration Methods: Results from a Large Scale Simulation Study. Econometric Reviews 29: 182-223. [CrossRef]
Zellner, Arnold, and Franz C. Palm. 1974. Time Series Analysis and Simultaneous Equation Econometric Models. Journal of Econometrics 2: 17-54. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Article

# Cointegration, Root Functions and Minimal Bases 

Massimo Franchi ${ }^{1, *}$ and Paolo Paruolo ${ }^{2}$<br>1 Department of Statistical Sciences, Sapienza University of Rome, P.le A. Moro 5, 00185 Rome, Italy<br>2 European Commission Joint Research Centre, Via E. Fermi 2749, 21027 Ispra, Italy; paolo.paruolo@ec.europa.eu<br>* Correspondence: massimo.franchi@uniroma1.it


#### Abstract

This paper discusses the notion of cointegrating space for linear processes integrated of any order. It first shows that the notions of (polynomial) cointegrating vectors and of root functions coincide. Second, it discusses how the cointegrating space can be defined (i) as a vector space of polynomial vectors over complex scalars, (ii) as a free module of polynomial vectors over scalar polynomials, or finally (iii) as a vector space of rational vectors over rational scalars. Third, it shows that a canonical set of root functions can be used as a basis of the various notions of cointegrating space. Fourth, it reviews results on how to reduce polynomial bases to minimal order-i.e., minimal bases. The application of these results to Vector AutoRegressive processes integrated of order 2 is found to imply the separation of polynomial cointegrating vectors from non-polynomial ones.


Keywords: VAR; cointegration; I(d); vector spaces

Citation: Franchi, Massimo, and Paolo Paruolo. 2021. Cointegration, Root Functions and Minimal Bases. Econometrics 9: 31. https://doi.org/ 10.3390/econometrics9030031

Academic Editor: Rocco Mosconi

Received: 9 December 2019
Accepted: 13 August 2021
Published: 17 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.


Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/).

## 1. Introduction

In their seminal paper, Engle and Granger (1987) introduced the notion of cointegration and of cointegrating (CI) rank for processes integrated of order 1, or I(1). They did this in the following way: ${ }^{1}$

DEFINITION: The components of the vector $x_{t}$, are said to be co-integrated of order $d, b$, denoted $x_{t} \sim C I(d, b)$, if (i) all components of $x_{t}$, are $I(d)$; (ii) there exists a vector $\beta(\neq 0)$ so that $z_{t}=\beta^{\prime} x_{t} \sim I(d-b), b>0$. The vector $\beta$ is called the co-integrating vector.
[...] If $x_{t}$ has $p$ components, then there may be more than one co-integrating vector $\beta$. It is clearly possible for several equilibrium relations to govern the joint behavior of the variables. In what follows, it will be assumed that there are exactly $r$ linearly independent co-integrating vectors, with $r \leq p-1$, which are gathered together into the $p \times r$ array $\beta$. By construction the rank of $\beta$ will be $r$ which will be called the "co-integrating rank" of $x_{t}$.

Engle and Granger (1987) did not define explicitly the notion of cointegrating space, but just the cointegrating rank, which corresponds to its dimension; explicit mention of the cointegrating space was first made in Johansen (1988).

The Granger representation theorem in Engle and Granger (1987) showed that the cointegration matrix $\beta$ needs to be orthogonal to the Moving Average (MA) impact matrix of $\Delta x_{t}$. More precisely, for $\Delta x_{t}=C(L) \varepsilon_{t}$, the MA impact matrix $C(1)$ has rank equal to $p-r$ and representation $C(1)=\beta_{\perp} a^{\prime}$, where $\beta_{\perp}$ is a basis of the orthogonal complement of the space spanned by the columns of $\beta$ and $a$ is full column rank.

Johansen $(1991,1992)$ stated the appropriate conditions under which the Granger representation theorem holds for $I(1)$ and $I(2)$ Vector AutoRegressive processes (VAR) $A(L) x_{t}=\varepsilon_{t}$, where the AR impact matrix $A(1)$ has rank equal to $r<p$ and rank factorization $A(1)=-\alpha \beta^{\prime}$, with $\alpha$ and $\beta$ of full column rank. He defined the cointegrating space as the vector space generated by the column vectors $\beta_{j}$ in $\beta$ over the field of real numbers $\mathbb{R}$.

Johansen (1991) noted that $\mathcal{B}=\operatorname{row}_{\mathbb{R}}\left(\beta^{\prime}\right)$ is uniquely defined ${ }^{2}$ by the rank factorization $A(1)=-\alpha \beta^{\prime}$, but the choice of basis $\beta^{\prime}$ is arbitrary, i.e., $\beta^{\prime}$ is not identified. Hypotheses that do not constrain $\mathcal{B}$ are hence untestable. He proposed likelihood ratio tests on $\mathcal{B}$ and described asymptotic properties of a just-identified version of $\beta^{\prime}$. Later Johansen (1995) discussed the choice of basis $\beta^{\prime}$ as an econometric identification problem of a system of simultaneous equations (SSE) of cointegrating relations describing the long-run equilibria in the process. He discussed identification using linear restrictions, along the lines of the classical identification problem of SSE studied in econometrics since the early days of the Cowles Commission.

The observation in Johansen (1988) that the cointegrating vectors formed a vector space $\mathcal{B}$ was an important breakthrough. For instance, it addressed the question: 'How many cointegrating vectors should one estimate in a given system of dimension $p$ ?'. A proper answer is in fact: A set of $r$ linearly independent vectors, spanning the cointegrating space $\mathcal{B}$, i.e., a basis of $\mathcal{B}$.

Similarly, when assuming that a set of $p$ interest rates is described by an $\mathrm{I}(1)$ process, the notion of cointegrating space $\mathcal{B}$ enables one to discuss questions like 'How should one test that all interest rates spreads are stationary?'. In fact, if all $\binom{p}{2}=p(p-1) / 2$ interest rates differentials were stationary, then one should have cointegrating rank $r=p-1$, which gives a first testable hypothesis on the cointegrating rank. Moreover there is no need to test all possible interest rates differentials to be stationary, but, if the cointegrating rank has been found to be $p-1$, one can test that the cointegrating space is spanned by any set of linearly independent $r$ contrasts between pairs of interest rates. If the cointegrating rank is found to be $0<r<p-1$, one may still want to test the restriction that the cointegrating space $\mathcal{B}$ is a subspace of the linear space spanned by all contrasts.

These questions, and many more, found clear answers thanks to the introduction of the notion of cointegrating space. The recognition that the set of cointegrating vectors forms a vector space was then instrumental to represent any cointegrating vector as a linear combination of the ones in a basis of the vector space.

The notion of cointegrating space, together with the complementary notion of attractor space, has been recently discussed in the context of functional time series for infinite dimensional Hilbert space valued AR processes with unit roots, see Beare et al. (2017), Beare and Seo (2020), Franchi and Paruolo (2020), and for infinite dimensional Banach space valued AR processes with unit roots, see Seo (2019).

For systems with variables integrated of order $d, I(d)$, with $d=2,3, \ldots$ Granger and Lee (1989) and Engle and Yoo (1991) introduced the related notions of multicointegration and polynomial cointegration; see also Engsted and Johansen (2000). However, no proper discussion of cointegrating spaces or of their corresponding bases has been proposed in the literature for higher order systems.

The present paper closes this gap, making use of classical concepts in local spectral theory, see Gohberg et al. (1993). A central role is played by canonical system of root functions, which have already been exploited in Franchi and Paruolo $(2011,2016)$ to characterize the inversion of a matrix function, and used in Franchi and Paruolo (2019) to derive the generalization of the Granger-Johansen representation theorem for $\mathrm{I}(d)$ processes.

In order to simplify exposition, this paper focuses on unit roots at a single point $z_{\omega}$, indexed by frequency $\omega$. When $\omega \notin\{0, \pi\}$, the resulting matrices are complex-valued, and the symbol $F$ is taken to indicate $\mathbb{C}$. For $\omega \in\{0, \pi\}, F$ is taken instead to indicate $\mathbb{R}$. Unit roots at distinct seasonal frequencies different from 0 have been considered e.g., in Hylleberg et al. (1990), Gregoir (1999), Johansen and Schaumburg (1998), Bauer and Wagner (2012). Several of these papers paired frequencies $\pm \omega$ when $\omega \notin\{0, \pi\}$ to obtain real coefficient matrices for Equilibrium Correction (EC) representations; in order to keep exposition as simple as possible, this is not attempted in the present paper.

To the best of the authors' knowledge, local spectral theory tools are employed here for the first time to discuss the definition of cointegrating space for $I(d)$ processes, $d>1$, and related bases. It is observed that several candidate cointegrating spaces exists, corre-
sponding to different choices of the set of vectors and scalars. The sets of vectors are chosen here to be either the set of polynomial vectors or the one of rational vectors, while the set of scalars are taken to be (i) the field $F=\mathbb{R}, \mathbb{C}$, (ii) the ring of polynomials with coefficients in $F$ (denoted $F[z]$ ) or (iii) the field of rational function with coefficients in $F$ (denoted $F(z)$ ). The resulting spaces are either vector spaces, in cases (i) and (iii), or a free module in case (ii). The relationship among their bases is discussed following Forney (1975), whose results are used to derive a polynomial basis of minimal degree-i.e., a minimal basis.

The focus of this paper is on the parsimonious representation of the set of cointegrating vectors. As noted by a referee, the present results may find application also in the parametrization and estimation of $I(d)$ EC systems. This, however, is beyond the scope of the present paper.

The rest of the paper is organised as follows. Section 2 provides the motivation for the paper. Section 3 reports definitions of integration and cointegration in $I(d)$ systems, where the cointegrating vectors $\zeta(z)^{\prime}=\sum_{j=0}^{\infty}\left(z-z_{\omega}\right)^{\prime} \zeta_{j}^{\prime}$ are allowed to be vector functions; here $\left(z-z_{\omega}\right)$ and its powers are associated with the difference operator and its powers. Section 4 defines root functions and canonical systems of root functions and Section 5 discusses possible definitions of the cointegration space. Section 6 discusses how to derive bases for the various notions of cointegrating space from VAR coefficients. Section 7 discusses minimal bases using results in Forney (1975) and Section 8 applies these results in order to obtain a minimal basis in the $I(2)$ VAR case. Section 9 concludes; Appendix A reports background results.

## 2. Motivation

This section motivates the study of the represention of cointegrating vectors in terms of bases of suitable spaces, for systems integrated of order two, which are more formally introduced in Section 3 below. Let $x_{t}$ be a $p \times 1$ vector process, and let $\Delta=1-L$ and $L$ be the (0-frequency) difference and the lag operators. Assume that $x_{t}$ is integrated of order 2, $\mathrm{I}(2)$, with $\Delta^{j} x_{t}$ nonstationary for $j<2$ and stationary for $j \geq 2$.

Mosconi and Paruolo (2017) consider the identification problem for the following cointegrating SSE with $I(2)$ variables

$$
\operatorname{ecm}_{t}=\xi(\Delta)^{\prime} x_{t}, \quad \text { with } \quad \xi(\Delta)^{\prime}:=\left(\begin{array}{c}
\beta^{\prime}+v^{\prime} \Delta \\
\gamma^{\prime} \Delta \\
\beta^{\prime} \Delta
\end{array}\right) \begin{aligned}
& \} r_{0} \\
& \} r_{1} \\
& \} r_{0}
\end{aligned}
$$

The first set of $r_{0}$ polynomial vectors has coefficient $\beta^{\prime}$ of order 0 (i.e., that multiplies $\Delta^{0}$ ) and coefficient $v^{\prime}$ of order 1 (i.e., that multiplies $\Delta^{1}$ ). The last $r_{0}+r_{1}$ polynomial vectors have 0 coefficients of order 0 and $\gamma^{\prime}$ and $\beta^{\prime}$ coefficients of order 1 . They discussed identification of the SSE with respect to transformations corresponding to pre-multiplication of $\xi(\Delta)^{\prime}$ (or ecm ${ }_{t}$ ) by a block triangular, nonsingular matrix of the form

$$
Q=\left(\begin{array}{ccc}
Q_{00} & Q_{0 \gamma} & Q_{0 \beta} \\
0 & Q_{\gamma \gamma} & Q_{\gamma \beta} \\
0 & 0 & Q_{00}
\end{array}\right)
$$

where $Q_{a b}$ are blocks of real coefficients, $a, b \in\{0, \gamma, \beta\}$, with $Q_{00}$ and $Q_{\gamma \gamma}$ nonsingular square matrices.

They show that $Q \xi(\Delta)^{\prime}=\xi^{\circ \prime}(\Delta)$ has the same structure as $\xi(\Delta)^{\prime}$ in terms of the null coefficient of order 0 in the last $r_{1}+r_{0}$ equations, as well as the same $\beta$ block as the coefficient of order 0 in the first $r_{0}$ and as the coefficient of order 1 in the last $r_{0}$ rows. More precisely,

- $\quad \beta^{\prime}$ is replaced by $\beta^{\circ \prime}=Q_{00} \beta^{\prime}$, a set of $r_{0}$ linear combinations of $\beta^{\prime}$,
- $\gamma^{\prime} \Delta$ is replaced by a set of $r_{1}$ linear combinations of $\gamma^{\prime} \Delta$ and $\beta^{\prime} \Delta$,
- $v^{\prime} \Delta$ is replaced by a set of $r_{0}$ linear combinations of $v^{\prime} \Delta, \gamma^{\prime} \Delta$ and $\beta^{\prime} \Delta$.

Remark 1 ( $F$-linear combinations). Note that the $Q$ linear combinations have scalars taken from $F=\mathbb{R}$, and that any CI vectors can be obtained as linear combinations with coefficients in $F$ of the rows in $\xi(\Delta)$ ', called in the following ' $F$-linear combinations'.

The main motivation to study the notion of cointegration space for $I(d)$ processes with $d \geq 2$ comes from the following observation.

Remark 2 ( $F[\Delta]$-linear combinations). The set of CI vectors obtained as F-linear combinations of the rows in $\xi(\Delta)^{\prime}$ can be also obtained by considering the alternative set of cointegrating vectors

$$
\left.\zeta(\Delta)^{\prime}:=\binom{\beta^{\prime}+v^{\prime} \Delta}{\gamma^{\prime}}\right\} \begin{aligned}
& r_{0} \\
& r_{1}
\end{aligned}
$$

and choosing linear combinations with scalar in the set of polynomials $F[\Delta]$, where $a(z) \in F[z]$ has the form $a(z)=\sum_{j=0}^{n} a_{j} z^{j}$ for some finite $n$.

To show that the set of $F[\Delta]$-linear combinations of $\zeta(\Delta)^{\prime}$ is the same as the set of $F$-linear combinations of $\xi(\Delta)^{\prime}$, it is sufficient to show that the rows of $\xi(\Delta)^{\prime}$ can be obtained as $F[\Delta]$-linear combinations of the rows in $\zeta(\Delta)^{\prime}$, possibly up to terms of the type $c^{\prime} \Delta^{2}$ which generate stationary processes by definition.

Note first that $\beta^{\prime}+v^{\prime} \Delta$ is common to $\xi(\Delta)^{\prime}$ and $\zeta(\Delta)^{\prime}$. In order to obtain $\gamma^{\prime} \Delta$ in $\xi(\Delta)^{\prime}$ from $\zeta(\Delta)^{\prime}$ one needs to select the scalar $\Delta$ from $F[\Delta]$ and multiply it by $\gamma^{\prime}$. Similarly, in order to obtain $\beta^{\prime} \Delta$ in $\xi(\Delta)^{\prime}$ one only needs to select the scalar $\Delta \in F[\Delta]$ and multiply it by $\beta^{\prime}+v^{\prime} \Delta$ to obtain $\beta^{\prime} \Delta+v^{\prime} \Delta^{2}$. Because $\Delta^{2} x_{t}$ is stationary by the assumption that $x_{t}$ is $I(2)$, the term $v^{\prime} \Delta^{2}$ can be discarded, and this completes the argument.

The take-away from Remark 2 is that, if one allows the set of multiplicative scalars to contain polynomials, i.e., if one moves from $F$-linear combinations to $F[\Delta]$-linear combinations, then one can reduce the number of rows needed to generate the set of CI vectors: $\xi(\Delta)^{\prime}$ in fact has $2 r_{0}+r_{1}$ rows, while the number of rows in $\zeta(\Delta)^{\prime}$ is $r_{0}+r_{1}$.

The previous discussion shows that the two sets, $F$ and $F[\Delta]$, could be used as possible set of scalars in taking linear combinations. The first one, $F$, is a field (i.e., a division ring), the second one, $F[\Delta]$, is a ring but not a field because it lacks the multiplicative inverse.

Given that vector spaces require the set of scalars to be a field, one may also consider another possible set of scalars, namely $F(\Delta)$, the set of rational functions of the type $a(\Delta)=c(\Delta) / d(\Delta)$ with $c(\Delta), d(\Delta) \in F[\Delta]$, and $d(\Delta)$ not identically equal to 0 , indicated as $d(\Delta) \not \equiv 0$. This leads to consider three possible choices for the set of scalars: (i) The field $F$, (ii) the ring $F[\Delta]$ and (iii) the field $F(\Delta)$. The rest of the paper discusses relative merits of using any of them.

The above discussion focused on unit roots at $z=1$, which are associated to the long run behavior of the process. When data are observed every month or quarter, seasonal unit roots, seasonal cointegration and seasonal error correction have been shown to be useful notions, see Hylleberg et al. (1990). For instance, in the case of quarterly series, the relevant seasonal unit roots are at $z=-1$ and at $z \pm i$ where $i$ is the imaginary unit. These roots are represented as $z_{\omega}=\exp (i \omega)$ with $0 \leq \omega<2 \pi$, where $z_{\omega}=1, i,-1,-i$ correspond to $\omega=0, \frac{1}{2} \pi, \pi, \frac{3}{2} \pi$.

Johansen and Schaumburg (1998) showed that the conditions under which a VAR process allows for seasonal integration (and cointegration) of order 1 are of the same type as for roots at $z=1$, except that expansions of the VAR polynomial are performed around each $z_{\omega}$, see their Theorem 3. They also provided the corresponding EC form in their Corollary 2; see also Bauer and Wagner (2012) and the discussion in Remark 9 below.

In general, the conditions for integration of any order $d$ at a point $z_{\omega}$ on the unit circle can be shown to be of the same type. This paper hence considers the generic case of a linear process with a generic root on the unit circle $z_{\omega}=\exp (i \omega)$, and discusses the notions of cointegration, root functions and minimal bases in this general context. This allows to show that the present results hold for generic frequency $\omega, 0 \leq \omega<2 \pi$.

Incidentally, the results presented below in Section 6 state the generalization of the Granger and the Johansen Representation Theorems presented in Franchi and Paruolo (2019) for a generic unit root $z_{\omega}=\exp (i \omega)$ at any frequency $\omega$.

## 3. Setup and Definitions

This section introduces notation and basic definitions of integrated and cointegrated processes.

### 3.1. Linear Processes

Assume that $\left\{\varepsilon_{t}, t \in \mathbb{Z}\right\}$ is a $p \times 1$ i.i.d. sequence, called a noise process, ${ }^{3}$ with $\mathrm{E}\left(\varepsilon_{t}\right)=0$ and $\mathrm{E}\left(\varepsilon_{t} \varepsilon_{s}^{\prime}\right)=\Omega 1_{s=t}$ where 1 . is the indicator function, and define the linear process $u_{t}=\mu_{t}+C(L) \varepsilon_{t}$, where $\mu_{t}$ is a nonstochastic $p \times 1$ vector and $C(z)=\sum_{j=0}^{\infty} z^{j} C_{j}^{\circ}$ is a $p \times p$ matrix function, with coefficient matrices $C_{j}^{\circ} \in \mathbb{R}^{p \times p}$. Note that the matrices $C_{j}^{\circ}$ are defined by an expansion of $C(z)$ around $z=0$. The term $\mu_{t}$ is nonstochastic, i.e., $\mathrm{E}\left(\mu_{t}\right)=\mu_{t}$, and can contain deterministic terms. Because $\mathrm{E}\left(\varepsilon_{t}\right)=0$, one sees that $\mathrm{E}\left(u_{t}\right)=\mu_{t}$, and hence in the following $u_{t}$ is often written as $u_{t}=\mathrm{E}\left(u_{t}\right)+C(L) \varepsilon_{t}$.

The matrix function $C(z)=\sum_{j=0}^{\infty} z^{j} C_{j}^{\circ}$ is assumed to be finite when $z$ is inside the open disk $D(0,1+\eta), \eta>0$, in $\mathbb{C}$ with center at 0 and radius $1+\eta>1$, i.e., $C(z)$ is assumed analytic on $D(0,1+\eta)$. Here and in the following $|\cdot|$ indicates the modulus and $D\left(z_{\star}, \rho\right)$ indicates the open disk $D\left(z_{\star}, \rho\right):=\left\{z \in \mathbb{C}:\left|z-z_{\star}\right|<\rho\right\}$ with center $z_{\star} \in \mathbb{C}$ and radius $\rho>0$. In this paper $C(z)$ is assumed to be regular on $D(0,1+\eta)$, i.e., $C(z)$ can lose rank only at a finite number of isolated points in $D(0,1+\eta)$.

Because of analyticity of $C(z)$, it can be expanded around any interior point of $D(0,1+\eta)$. In particular, define the point $z_{\omega}:=e^{i \omega}$ on the unit circle at frequency $\omega, \omega \in[0,2 \pi)$, and observe that it lies inside $D(0,1+\eta)$ because $\eta>0$. Hence one can expand $C(z)$ as $C(z)=\sum_{j=0}^{\infty}\left(z-z_{\omega}\right)^{j} C_{j}$ on $D\left(z_{\omega}, \eta\right), \eta>0$. Note that the matrices $C_{j}$ are defined by an expansion of $C(z)$ around $z=z_{\omega}$, but that the dependence of $C_{j}$ on $\omega$ is not included in the notation for simplicity. The analysis of the properties of $C(z)$ is done locally around $z=z_{\omega}$ on $D\left(z_{\omega}, \eta\right), \eta>0$.

Similarly to $C(z)$, one can consider a scalar function of $z, a(z)$ say, or a $1 \times p$ vector function $b(z)^{\prime}$ taken to be analytic on $D\left(z_{\omega}, \eta\right), \eta>0$. This means that $a(z)$ has representation $a(z)=\sum_{j=0}^{\infty}\left(z-z_{\omega}\right)^{j} a_{j}$ around $z_{\omega}$ and similarly for $b(z)^{\prime}$. A special case is when $a(z)$ is a polynomial of degree $k, a(z)=\sum_{j=0}^{k}\left(z-z_{\omega}\right)^{j} a_{j}$, which corresponds to setting all $a_{j}=0$ for $j>k$. Another special case is given by rational functions $a(z)=c(z) / d(z)$ with $c(z)$ and $d(z)$ polynomials, where $d(z) \not \equiv 0$ and $z_{\omega}$ is not a root of $d(z)$. Similarly for $b(z)^{\prime}$.

### 3.2. Integration

The following definition specifies the $I_{\omega}(0)$ class of processes as a subset of all linear processes built from the noise sequence $\varepsilon_{t}$, and introduces the notion of $I_{\omega}(d)$ processes using the difference operator at frequency $\omega, \Delta_{\omega}:=1-e^{-i \omega} L=1-z_{\omega}^{-1} L$. To simplify notation, the dependence of $\Delta_{\omega}$ on the lag operator $L$ is left implicit. Observe also that, because $z_{\omega}=e^{i \omega} \neq 0, z-z_{\omega}$ in the analytic expansions can be expressed as $\left(-z_{\omega}\right)(1-$ $\left.z / z_{\omega}\right)$, where $\left(1-z / z_{\omega}\right)$ corresponds to the operator $\Delta_{\omega}$.

Next, the definition of order of integration is introduced; this is defined as the difference between two nonnegative integer exponents $d_{1}$ and $d_{2}$ of $\Delta_{\omega}$ in the representation that links the process $x_{t}$ with its driving linear process $u_{t}$. This definition allows for the possibility to have $x_{t}$ integrated of negative order.

Definition 1 (Integrated processes at frequency $\omega$ ). Let $C(z)$ be analytic on $D(0,1+\eta)$, $\eta>0$, and let $\varepsilon_{t}$ be a noise process. If $\left\{u_{t}, t \in \mathbb{Z}\right\}$, satisfies $u_{t}=\mathrm{E}\left(u_{t}\right)+C(L) \varepsilon_{t}$, then $u_{t}$ is called a linear process; if, in addition,

$$
\begin{equation*}
C\left(z_{\omega}\right) \neq 0, \quad z_{\omega}=\exp (i \omega), \quad 0 \leq \omega<2 \pi \tag{1}
\end{equation*}
$$

then $u_{t}$ is said to be integrated of order zero at frequency $\omega$, indicated $u_{t} \sim I_{\omega}(0)$.
Let $d_{1}, d_{2} \in \mathbb{N}_{0}=\mathbb{N} \cup 0$ be finite non-negative integers; if $\left\{x_{t}, t \in \mathbb{Z}\right\}$ satisfies $\Delta_{\omega}^{d_{1}}\left(x_{t}-\right.$ $\left.\mathrm{E}\left(x_{t}\right)\right)=\Delta_{\omega}^{d_{2}}\left(u_{t}-\mathrm{E}\left(u_{t}\right)\right)$ where $u_{t} \sim I_{\omega}(0)$, then $x_{t}$ is said to be integrated of order $d:=d_{1}-d_{2}$ at frequency $\omega$, indicated $x_{t} \sim I_{\omega}(d)$; in this case $x_{t}$ has representation

$$
\begin{equation*}
\Delta_{\omega}^{d_{1}}\left(x_{t}-\mathrm{E}\left(x_{t}\right)\right)=\Delta_{\omega}^{d_{2}} C(L) \varepsilon_{t} \tag{2}
\end{equation*}
$$

where $C(z)$ is analytic on $D(0,1+\eta), \eta>0$, and $C\left(z_{\omega}\right) \neq 0$.
Remark 3 (Negative orders). When $d_{1}<d_{2}$, the integration order $d:=d_{1}-d_{2}$ is negative. Note also that Definition 1 avoids to define the operator $\Delta_{\omega}^{-1}$; see however Equations (5) and (6) below.

Remark 4 (Mean-0 linear process). The linear process $u_{t}$ in Definition 1 can have any expectation $\mathrm{E}\left(u_{t}\right)$, which however, does not play any role in the definition of the $x_{t}$ process. Hence, one can assume that $\mathrm{E}\left(u_{t}\right)=0$ in Definition 1 without loss of generality.

Remark 5 ( $\mathrm{E}\left(x_{t}\right)$ in Definition 1). Assume $x_{t}=\cos (2 t)+\exp (-3 t)+C(L) \varepsilon_{t}$ with $C(z)$ analytic on $D(0,1+\eta), \eta>0$, and $C\left(z_{\omega}\right) \neq 0, z_{\omega}:=e^{i \omega}$. Then $\mathrm{E}\left(x_{t}\right)=\cos (2 t)+\exp (-3 t)$ and Definition 1 implies that $x_{t}$ is $I_{\omega}(0)$. This example shows that the presence of $\mathrm{E}\left(x_{t}\right)$ in Equation (2) allows to concentrate attention on the stochastic part of the process $x_{t}$.

Remark 6 (Preference for low $d_{1}, d_{2}$ ). Assume for instance that (2) is satisfied for $\left(d_{1}, d_{2}\right)=$ $(1,0)$, and observe that this implies that (2) is satisfied for $\left(d_{1}, d_{2}\right)=(1+m, m)$ for any $m \in \mathbb{N}$. In the following, preference is given to the minimal pair $\left(d_{1}, d_{2}\right)$ for which (2) is satisfied, i.e., to $\left(d_{1}, d_{2}\right)=(1,0)$ in the example.

Leading cases are the ones where either $d_{1}$ or $d_{2}$ equals 0 . Specifically, when $0=d_{1}<d_{2}$, $d=d_{1}-d_{2}=-d_{2}$ is negative, and (2) reads

$$
\begin{equation*}
x_{t}-\mathrm{E}\left(x_{t}\right)=\Delta_{\omega}^{d_{2}} C(L) \varepsilon_{t} . \tag{3}
\end{equation*}
$$

When $d_{1} \geq d_{2}=0$ and hence $d=d_{1}-d_{2}=d_{1}$ is nonnegative, (2) reads

$$
\begin{equation*}
\Delta_{\omega}^{d_{1}}\left(x_{t}-\mathrm{E}\left(x_{t}\right)\right)=C(L) \varepsilon_{t} \tag{4}
\end{equation*}
$$

Remark 7 (Example of $I_{0}(-1)$ ). As an example, consider the process $x_{t}=C(L) \varepsilon_{t}$ with $C(L)=$ $1-$ L. Setting $\omega=0$ one finds that Equation (2) is satisfied with $d=-1$, i.e., that the process is $I_{0}(-1)$. Selecting any other frequency $0<\omega<2 \pi$, one sees that Equation (2) is satisfied for $d=0$, i.e., that the order of integration is 0 , i.e., $I_{\omega}(0)$ for $0<\omega<2 \pi$. This illustrates the fact that a process may have different orders of integration at different frequencies.

Remark $8\left(t \in \mathbb{Z}\right.$ versus $\left.t \in \mathbb{N}_{0}\right)$. Consider the process $x_{t}=c+\sum_{j=1}^{t} \varepsilon_{t}$ defined only for $t \in \mathbb{N}_{0}=\mathbb{N} \cup 0$, which satisfies $\Delta_{0}\left(x_{t}-c\right)=\varepsilon_{t}$ for $t \in \mathbb{N}$. Consider another process $\left\{x_{t}^{\star}, t \in \mathbb{Z}\right\}$ satisfying the same equation $\Delta_{0}\left(x_{t}^{\star}-c\right)=\varepsilon_{t}$ for $t \in \mathbb{Z}$ with $x_{t}=x_{t}^{\star}$ for $t \in \mathbb{N}_{0}$. The process $\left\{x_{t}^{\star}, t \in \mathbb{Z}\right\}$ is $I_{0}(1)$ according to Definition 1 , and it is suggested to extend this qualification to $x_{t}$, because it coincides with the $x_{t}^{\star}$ process on the non-negative integers, $x_{t}=x_{t}^{\star}$ for $t \in \mathbb{N}_{0}$.

Remark 9 (One or more frequencies). Definition 1 of integration refers to a single frequency $\omega$, but it can be used to cover multiple frequencies. In fact, consider the 'ARMA process with unit root structure', as defined in Bauer and Wagner (2012), i.e., a process $x_{t}$ satisfying $D(L) x_{t}=v_{t}$ where $D(L):=\prod_{j=1}^{n} \Delta_{\omega_{j}}^{m_{i}}$ for a (finite) set of frequencies $\omega_{1}, \ldots, \omega_{n}$, with $v_{t}$ a stationary ARMA process $v_{t}=C(L) \varepsilon_{t}$ with $C\left(\exp \left(i \omega_{j}\right)\right) \neq 0$. They call $\left\{\left(\omega_{j}, m_{j}\right), j=1, \ldots, n\right\}$, the 'unit root structure' of $x_{t}$, see their Definition 2. This can be obtained using Definition 1 for each $\omega_{j}$ in turn, noting that $v_{t}$ being ARMA corresponds to a rational $C(z)$, which is a special case of the definition above.

Hylleberg et al. (1990), Gregoir (1999), Johansen and Schaumburg (1998), Bauer and Wagner (2012) consider $x_{t}$ to be real-valued, which implies that integration frequencies $\pm \omega_{j}$ are 'paired', so that if $\exp \left(i \omega_{j}\right)$ is a unit root of the process, so is $\exp \left(-i \omega_{j}\right)$; this implies that in this case one can pair frequencies $\pm \omega_{j}$ with $0<\omega_{j}<\pi$ and rearrange coefficients so as to obtain real coefficient matrices in EC representations. This is not done in this paper for reasons of simplicity.

Remark 10 (Relation with other definitions). The definition of an $I_{\omega}(0)$ (respectively an $I_{\omega}(d)$ ) process in the present Definition 1 coincides with Definition 3.2 (respectively Definition 3.3) in Johansen (1996) when setting $\omega=0$ (respectively $\omega=0$ and $d_{2}=0$ ). The present definition also agrees with Definitions 2.1 and 2.2 of integration in Gregoir (1999), both for positive and negative orders and any frequency $\omega$. The definition also agrees with the one in Franchi and Paruolo (2019) when applied to vector processes.

Remark 11 (Entries in $C(z)$ ). When $\omega$ differs from 0 or $\pi$, the point $z_{\omega}=e^{i \omega}$ has a nonzero complex part; hence the matrix $C\left(z_{\omega}\right)$ in (1) has complex entries and the coefficient matrices $C_{j}$ in the expansion $C(z)=\sum_{j=0}^{\infty}\left(z-z_{\omega}\right)^{j} C_{j}$ are complex even when the coefficients in the expansion around $z=0$ are real.

Following Gregoir (1999), the summation operator at frequency $\omega$ is defined as

$$
\begin{equation*}
\mathcal{S}_{\omega} u_{t}:=1_{t>0} \sum_{j=1}^{t} u_{j} e^{-i \omega(t-j)}-1_{t<0} \sum_{j=t+1}^{0} u_{j} e^{-i \omega(t-j)} \tag{5}
\end{equation*}
$$

Basic properties of the operator are proved in Gregoir (1999); these include

$$
\begin{equation*}
\Delta_{\omega} \mathcal{S}_{\omega} u_{t}=u_{t,} \quad \mathcal{S}_{\omega} \Delta_{\omega} u_{t}=u_{t}-u_{0} e^{-i \omega t} \tag{6}
\end{equation*}
$$

where $\left\{u_{t}, t \in \mathbb{Z}\right\}$ is any sequence over $\mathbb{Z}$.
Remark 12 (Simplifications of $\Delta_{\omega}$ and initial values). Take $d_{1}=d_{2}=1$ in (2), which in this case reads $\Delta_{\omega} x_{t}=\Delta_{\omega} u_{t}$ with $u_{t} \sim I_{\omega}(0)$. Applying the $\mathcal{S}_{\omega}$ operator on both sides one obtains $x_{t}-x_{0} e^{-i \omega t}=u_{t}-u_{0} e^{-i \omega t} .^{4}$ If one assigns the initial value of $x_{0}$ equal to $u_{0}$, one obtains $x_{t}=u_{t}$, which corresponds to the cancellation of $\Delta_{\omega}$ from both sides of (2). The same reasoning applies for generic $d_{1}, d_{2}>0$ to the cancellation of $\Delta_{\omega}^{\min \left(d_{1}, d_{2}\right)}$ from both sides of (2). This shows that one can simplify powers of $\Delta_{\omega}$ from both sides of (2) by properly assigning initial values; this cancellation is always implicitly performed in the following, in line with preference for minimal values of $d_{1}, d_{2}$ as discussed in Remark 6.

### 3.3. Cointegration

Cointegration is the property of (possibly polynomial) linear combinations of $x_{t}$ to have a lower order of integration with respect to the original order of integration of $x_{t}$ at frequency $\omega$. Specifically, consider a nonzero $1 \times p$ row vector function $\zeta(z)^{\prime}=$ $\sum_{j=0}^{\infty} \zeta_{j}^{\prime}\left(z-z_{\omega}\right)^{j}$, analytic on a disk $D\left(z_{\omega}, \eta\right), \eta>0$. As in Engle and Granger (1987), the idea is to call $\zeta(L)^{\prime}$ cointegrating if $\zeta(L)^{\prime} x_{t}$ has lower order of integration than $x_{t}$, excluding cases such as $\zeta(L)^{\prime}=\Delta_{\omega} a^{\prime}$ where $a^{\prime}$ by itself does not reduce the order of integration.

This leads to the following definition.
Definition 2 (Cointegrating vector at frequency $\omega$ ). Let $x_{t} \sim I_{\omega}(d)$ be as in Definition 1, i.e.,

$$
\Delta_{\omega}^{d_{1}}\left(x_{t}-\mathrm{E}\left(x_{t}\right)\right)=\Delta_{\omega}^{d_{2}} C(L) \varepsilon_{t}
$$

where $d:=d_{1}-d_{2}, C(z)$ is analytic on $D(0,1+\eta), \eta>0$, and $C\left(z_{\omega}\right) \neq 0$, see (2); let also $\zeta(z)^{\prime}=\sum_{j=0}^{\infty}\left(z-z_{\omega}\right)^{j} \zeta_{j}^{\prime}$ be a $1 \times p$ row vector function, analytic on $D\left(z_{\omega}, \eta\right)$ with $\zeta\left(z_{\omega}\right)^{\prime}=$
$\zeta_{0}^{\prime} \neq 0^{\prime}$. Then $\zeta(z)^{\prime}$ is called a cointegrating vector at frequency $\omega$ if $\zeta(L)^{\prime} x_{t} \sim I_{\omega}(d-s)$ for some $s \in \mathbb{N}$, i.e.,

$$
\begin{equation*}
\zeta(L)^{\prime} \Delta_{\omega}^{d_{1}}\left(x_{t}-\mathrm{E}\left(x_{t}\right)\right)=\Delta_{\omega}^{d_{2}+s} g(L)^{\prime} \varepsilon_{t} \tag{7}
\end{equation*}
$$

where $g(z)^{\prime}$ is analytic on $D\left(z_{\omega}, \eta\right), \eta>0$, and $g\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$. Given Equation (2), Equation (7) is equivalent to the condition

$$
\begin{equation*}
\zeta(L)^{\prime} C(L)=\Delta_{\omega}^{s} g(L)^{\prime}, \quad g\left(z_{\omega}\right)^{\prime} \neq 0^{\prime} \tag{8}
\end{equation*}
$$

The positive integer $s \in \mathbb{N}$ is called the order of the cointegrating vector $\zeta(z)^{\prime}$ of $C(z)$ at $z_{\omega}$. $x_{t}$ is said to be cointegrated at frequency $\omega$ if any cointegrating vector $\zeta(z)^{\prime}=\sum_{j=0}^{\infty}\left(z-z_{\omega}\right)^{j} \zeta_{j}^{\prime}$ can be replaced by $\zeta\left(z_{\omega}\right)^{\prime}=\zeta_{0}^{\prime}$ without decreasing the order $\sin (8)$; otherwise $x_{t}$ is said to be multicointegrated at frequency $\omega$.

Remark $13\left(C(z)\right.$ has full rank on $D\left(z_{\omega}, \eta\right), \eta>0$, except at $\left.z=z_{\omega}\right)$. Because cointegrating vectors are by definition different from zero at $z_{\omega}, x_{t}$ is cointegrated at frequency $\omega$ if and only if $C\left(z_{\omega}\right) \neq 0$ has reduced rank. Moreover, because $C(z)$ is regular on $D(0,1+\eta)$, the point $z_{\omega}$ is isolated, i.e., $C(z)$ has full rank on $D\left(z_{\omega}, \eta\right), \eta>0$, except at $z=z_{\omega}$.

Remark 14 (Entries in cointegrating vectors). Similarly to Remark 11, the coefficient vectors $\zeta_{j}^{\prime}$ in the expansion $\zeta(z)^{\prime}=\sum_{j=0}^{\infty}\left(z-z_{\omega}\right)^{j} \zeta_{j}^{\prime}$ are in general complex. Note that $\zeta(L)^{\prime}=\Delta_{\omega} a^{\prime}$ does not satisfy the definition because the requirement $\zeta\left(z_{\omega}\right)^{\prime}=\zeta_{0}^{\prime} \neq 0^{\prime}$ is not satisfied.

Remark 15 (d and s). Recall that $d$ (the order of integration) is the difference between the exponents of $\Delta_{\omega}$ on the l.h.s. and r.h.s. of (2). When pre-multiplied by $\zeta(L)^{\prime}$, the exponent on the r.h.s. decreases by s and the difference of the exponents on the l.h.s. and r.h.s. of (7) becomes $d-s$. Because $\zeta_{0}^{\prime} \neq 0^{\prime}$, this can only happen if $\zeta(L)^{\prime}$ factors $\Delta_{\omega}^{s}$ from $C(L)$, see (8). The condition $g\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$ guarantees that no remaining additional power of $\Delta_{\omega}$ can be factored from $C(L)$ using $\zeta(L)^{\prime}$.

Remark 16 (Examples of cointegration vectors). Take $\zeta(L)^{\prime}=\zeta_{0}^{\prime}$ with $\zeta_{0}$ chosen in $(\operatorname{col} C(1))^{\perp}$, and note that this implies $s \geq 1$ in (7). This shows that the definition contains the $I_{0}(1)$ definition of cointegrating vectors as a special case.

The usual definition of cointegration, see Definition 3.4 in Johansen (1996), considers a $p \times 1$ process $x_{t} \sim I_{0}(1)$ and defines $x_{t}$ cointegrated with cointegrating vector $\zeta \neq 0$ if $\zeta^{\prime} x_{t}$ "can be made stationary by a suitable choice of initial distribution". The following proposition clarifies that his definition coincides with the one in this paper.

Proposition 1 (Relation with Definition 3.4 in Johansen (1996)). $\zeta^{\prime}$ is a cointegrating vector in the sense of Definition 3.4 in Johansen (1996) if and only if Definition 2 is satisfied with $\omega=0$ and $\zeta(z)^{\prime}=\zeta^{\prime}, d=1, s \in \mathbb{N}$.

Proof. For simplicity and without loss of generality, set $\mathrm{E}\left(x_{t}\right)=0$ and omit the subscript $\omega=0$. Assume Definition 2 is satisfied with $\omega=0$ and $\zeta(z)^{\prime}=\zeta^{\prime}, d=1$, and $s \in \mathbb{N}$, i.e.,

$$
\begin{equation*}
\Delta \zeta^{\prime} x_{t}=\Delta^{s} g(L) \varepsilon_{t} \tag{9}
\end{equation*}
$$

see Remark 12, and set $v_{t}:=\Delta^{s-1} g(L) \varepsilon_{t}$. Applying $\mathcal{S}$ to both sides of Equation (9) one finds $\zeta^{\prime} x_{t}-\zeta^{\prime} x_{0}=v_{t}-v_{0}$. Note that $v_{t}$ is stationary for any $s \in \mathbb{N}$, and hence the initial values $\zeta^{\prime} x_{0}$ can be chosen equal to $v_{0}$, so as to obtain $\zeta^{\prime} x_{t}=v_{t}$, a stationary process.

Conversely, assume that $\zeta^{\prime}$ is a cointegrating vector in the sense of Definition 3.4 in Johansen (1996). Because $x_{t} \sim I(1)$, one has $\Delta x_{t}=C(L) \varepsilon_{t}$, see Definition 1, with $C(z)$ analytic on a disk $D(0,1+\eta), \eta>0$, which admits expansion $C(z)=C+\widetilde{C}(z)(1-z)$ around 1 , where $\widetilde{C}(z)$ is analytic on the same disc. A necessary and sufficient condition for
cointegration in the sense of Definition 3.4 in Johansen (1996) is that $\zeta^{\prime} C=0$ as shown in Johansen (1988) Equation (17); see also Engle and Granger (1987, p. 256). ${ }^{5}$ Hence one finds $\zeta^{\prime} \Delta x_{t}=\Delta g(L)^{\prime} \varepsilon_{t}$ with $g(z):=\zeta^{\prime} \widetilde{C}(z)$, which is analytic on $D(0,1+\eta), \eta>0$, and hence also on $D(1, \eta), \eta>0$. By Corollary 1 below, one has that $g(z)^{\prime}$ satisfies $g(z)^{\prime}=\Delta^{m} \widetilde{g}(z)^{\prime}$ with finite $m \in \mathbb{N}_{0}$ and $\widetilde{g}\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$. This shows that Definition 2 is satisfied with $\zeta(z)^{\prime}=\zeta^{\prime}$, $d=1$, and $s=m+1 \in \mathbb{N}$.

Remark 17 ( $\zeta^{\prime} x_{t}$ can have negative order of integration). Johansen (1996) makes the following observation just after his Definition 3.4: "Note that $\zeta^{\prime} x_{t}$ need not be $I(0)$ ", which recognises that $\zeta^{\prime} x_{t}$ can have negative order of integration. This is indeed the case when to $s=2,3, \ldots$ in Definition 2, because $\zeta^{\prime} x_{t} \sim I(1-s)$.

Remark 18 (Relation to other definitions in the literature). The definition of cointegration in Engle and Granger (1987) reported in the introduction is a special case of the present one with $\zeta(z)^{\prime}=\zeta_{0}^{\prime}$ a constant vector and $\omega=0$, under the additional requirement that all variables are integrated of the same order. For more details on this for the case $\omega=0$, see Franchi and Paruolo (2019). When $s>1$ and $\omega=0$, Definition 2 covers the definitions of multicointegration and polynomial cointegration in Granger and Lee (1989), Engle and Yoo (1991), Johansen (1996). When $s=1$ and $\omega=2 \pi j / n$ for $j=1, \ldots, n$ where $n$ is the number of seasons, the definition covers seasonal cointegration in Hylleberg et al. (1990), Johansen and Schaumburg (1998).

Example 1 (I(1) VAR). Following Johansen (1988), consider $A(L) x_{t}=\varepsilon_{t}$ with $A(z)=I-$ $\sum_{j=1}^{k}(1-z)^{j} A_{j}$ analytic on $\mathbb{C}$. Assume also that $\operatorname{det} A(z)=0$ has only solutions outside $D(0,1+$ $\eta), \eta>0$, or at $z=1$, where 'det' indicates the determinant of a matrix. Here and in the following, let $a_{\perp}$ indicate a basis of the orthogonal complement of the linear space spanned by the columns of the matrix $a$. Moreover $P_{a}:=a\left(a^{\prime} a\right)^{-1} a^{\prime}$ for a full-column-rank matrix $a$ is the orthogonal projection matrix onto $\operatorname{col}(a)$. Johansen (1991) (see his Equations (4.3) and (4.4) in Theorem 4.1) showed that for $x_{t}$ to be $I(1)$ at frequency $\omega=0$, a set of necessary and sufficient conditions are:
(i) $A(1)=-\alpha_{0} \beta_{0}^{\prime}$ with $\alpha_{0}, \beta_{0}$ full column rank matrices of dimension $p \times r_{0}, r_{0}<p$,
(ii) $P_{\alpha_{0 \perp}} A_{1} P_{\beta_{0 \perp}}=-\alpha_{1} \beta_{1}^{\prime}$ of maximal rank $r_{1}=p-r_{0}$.

In this case $x_{t}$ satisfies (2) for $d_{1}=1, d_{2}=0$, and $\zeta(L)^{\prime}=\zeta^{\prime}$ taken to be any row vector in $\mathcal{B}=\operatorname{row}_{F}\left(\beta_{0}^{\prime}\right)$ with $F=\mathbb{R}$.

Example 2 (I(2) VAR). Following Johansen (1992), consider the same VAR process as in Example 1. Johansen (1992) showed that for $x_{t}$ to be I(2) at frequency $\omega=0$, a set of necessary and sufficient conditions are:
(i) $A(1)=-\alpha_{0} \beta_{0}^{\prime}$ with $\alpha_{0}, \beta_{0}$ full column rank matrices of dimension $p \times r_{0}, r_{0}<p$,
(ii) $P_{\alpha_{0 \perp}} A_{1} P_{\beta_{0 \perp}}=-\alpha_{1} \beta_{1}^{\prime}$ with $\alpha_{1}, \beta_{1}$ full column rank matrices of dimension $p \times r_{1}, r_{1}<p-r_{0}$,
(iii) $P_{\left(\alpha_{0}, \alpha_{1}\right)_{\perp}}\left(A_{2}+A_{1} \bar{\beta}_{0} \bar{\alpha}_{0}^{\prime} A_{1}\right) P_{\left(\beta_{0}, \beta_{1}\right)_{\perp}}=-\alpha_{2} \beta_{2}^{\prime}$ of maximal rank $r_{2}=p-r_{0}-r_{1}$.

In this case $x_{t}$ satisfies (2) for $d_{1}=2, d_{2}=0$, and $\zeta(L)^{\prime}=\zeta_{0}^{\prime}+\Delta \zeta_{1}^{\prime}$ taken to be any row vector obtained by linear combinations of the rows in $\beta_{0}^{\prime}+(1-L) \bar{\alpha}_{0}^{\prime} A_{1}$ and $\beta_{1}^{\prime}$. The notion of cointegrating space for $I(2)$ processes is discussed in detail below, where $\bar{\alpha}_{0}^{\prime} A_{1}$ is called the 'multicointegrating coefficient'.

## 4. Root Functions, Cointegrating Vectors and Canonical Systems

This section introduces root functions and canonical systems of root functions, and their connection to cointegrating vectors, as defined in Definition 2 above.

### 4.1. Root Functions

Let $x_{t} \sim I_{\omega}(d)$ be cointegrated at frequency $\omega$, i.e., see Definition 2,

$$
\Delta_{\omega}^{d_{1}}\left(x_{t}-\mathrm{E}\left(x_{t}\right)\right)=\Delta_{\omega}^{d_{2}} C(L) \varepsilon_{t}
$$

where $d:=d_{1}-d_{2}$ and $C(z)$ has full rank on $D\left(z_{\omega}, \eta\right), \eta>0$, except at $z=z_{\omega}$, see Remark 13.

The following definition of (left) root functions is taken from Gohberg et al. (1993); this definition is given in a neighborhood of $z_{\omega}$.

Definition 3 (Root function). A $1 \times p$ row vector function $\varphi(z)^{\prime}$ analytic on $D\left(z_{\omega}, \eta\right)$ is called a root function of $C(z)$ at $z_{\omega}$ if $\varphi\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$ and if

$$
\begin{equation*}
\varphi(z)^{\prime} C(z)=\left(z-z_{\omega}\right)^{s} \widetilde{\varphi}(z)^{\prime}, \quad s \in \mathbb{N}, \quad \widetilde{\varphi}\left(z_{\omega}\right)^{\prime} \neq 0^{\prime} \tag{10}
\end{equation*}
$$

The positive integer s is called the order of the root function $\varphi(z)^{\prime}$ at $z_{\omega}$.
Observe that $\widetilde{\varphi}(z)^{\prime}$ is $1 \times p$ and analytic on $D\left(z_{\omega}, \eta\right), \eta>0$.
Remark 19 (Factoring the difference operator). Definition 3 characterizes roots functions by their ability to factor powers of $\left(z-z_{\omega}\right)$ from $C(z)$. Note that, because here $z_{\omega}=\exp (i \omega) \neq 0$, one can write $\left(z-z_{\omega}\right)$ as $\left(-z_{\omega}\right)\left(1-z / z_{\omega}\right)$ where $\left(1-z / z_{\omega}\right)$ corresponds to the difference operator $\Delta_{\omega}$ and $\left(-z_{\omega}\right)$ can be absorbed in $\widetilde{\varphi}(z)^{\prime}$ without affecting its property that $\widetilde{\varphi}\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$.

Remark 20 (Local analysis). Note first that $C(z)$ cannot be identically 0 in Definition 3, because $C(z)$ is assumed to be regular. Next take for example the $2 \times 2$ matrix $C(z)=\operatorname{diag}((1-z),(1+z))$ which has full rank on $\mathbb{C}$, except at the two points $z_{0}=1$ and $z_{\pi}=-1$, where it has rank 1 .

Take first the point at $z_{0}=1$; in this case one could choose a disk $D(1, \eta)$ with any $\eta<2$, on which $C(z)$ is analytic and full rank except at $z_{0}=1$. One can verify that a root function is $\varphi_{1}(z)^{\prime}=(1,0)$, which satisfies $\varphi_{1}(z)^{\prime} C(z)=(1-z) \widetilde{\varphi}_{1}(z)^{\prime}$ with $\widetilde{\varphi}_{1}(z)^{\prime}=(1,0)$. The same can be repeated for the other point $z_{\pi}=-1$, choosing a different disk $D(-1, \eta)$ with any $\eta<2$, and a root function equal to $(0,1)$.

The implication of this example is that one can have multiple separated points where $C(z)$ has reduced rank, and apply the above definition to each point separately, using a different disk $D$ for each point. In other words, the discussion of cointegration in this paper is local to a single unit root.

Remark 21 (Order). A root function factorises $\left(z-z_{\omega}\right)^{s}$ from $C(z)$, and s indicates the order. The condition $\varphi\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$ guarantees that in the analytic expansion $\varphi(z)^{\prime}=\sum_{n=0}^{\infty}\left(z-z_{\omega}\right)^{n} \varphi_{n}^{\prime}$, the first term $\varphi_{0}^{\prime}$ is not the null vector. Note that the condition $\widetilde{\varphi}\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$ makes sure that one cannot extract additional factors of $\left(z-z_{\omega}\right)$ from $C(z)$ using $\varphi(z)^{\prime}$.

It is immediate to see that a cointegrating vector is a root function of $C(z)$ and vice versa, as stated in the following theorem.

Theorem 1 (Cointegrating vectors and root functions). $\zeta(z)^{\prime}$ is a cointegrating vector at frequency $\omega$ if and only if $\zeta(z)^{\prime}$ is a root function of $C(z)$ at $z_{\omega}$, and the order of the cointegrating vector and of the root function coincide.

Proof. Observe that any root function satisfies Definition 2 of cointegrating vectors and vice versa, including the definition of their order.

Results in Gohberg et al. (1993) shows that the order of a root functions is finite, because it is bounded by the order of $z_{\omega}$ as a zero of $\operatorname{det} C(z)$, a result that is reported in the next proposition.

Proposition 2 (Bound on the order of a root function). The order of a root function of $C(z)$ at $z_{\omega}$ is at most equal to the order of $z_{\omega}$ as a zero of $\operatorname{det} C(z)$, which is finite because $C(z)$ is regular.

Proof. See Gohberg et al. (1993).

Corollary 1 (Bound on the order of a cointegrating vector). The order of any cointegrating vector at frequency $\omega$ is finite.

Proof. This follows from Proposition 2 because cointegrating vectors and root functions coincide by Theorem 1.

### 4.2. Canonical Systems of Root Functions

Next, canonical systems of root functions for $C(z)$ at $z_{\omega}$ are introduced, see Gohberg et al. (1993). Choose a root function $\phi_{1}(z)^{\prime}$ of highest order $s_{1}$. Since the orders of the root functions are bounded by Proposition 2, such a function exists. Next proceed iteratively over $j=2, \ldots$, choosing the next root function $\phi_{j}(z)^{\prime}$ to be of the highest order $s_{j}$ such that $\phi_{j}\left(z_{\omega}\right)^{\prime}$ is linearly independent from $\phi_{1}\left(z_{\omega}\right)^{\prime}, \ldots, \phi_{j-1}\left(z_{\omega}\right)^{\prime}$. Because $m:=\operatorname{dim}\left(\left(\operatorname{col} C\left(z_{\omega}\right)\right)^{\perp}\right)<\infty$, this process ends with $m$ root functions $\phi_{1}(z)^{\prime}, \ldots, \phi_{m}(z)^{\prime}$.

Note that the columns in $a:=\left(\phi_{1}\left(z_{\omega}\right), \ldots, \phi_{m}\left(z_{\omega}\right)\right)$ span the finite dimensional space $\left(\operatorname{colC}\left(z_{\omega}\right)\right)^{\perp}$, so that one can choose vectors $\left(\phi_{m+1}, \ldots, \phi_{p}\right)=a_{\perp}$ that span its orthogonal complement. This construction leads to the following definition.

Definition 4 ((Extended) canonical system of root functions). Let $\phi_{1}(z)^{\prime}, \ldots, \phi_{m}(z)^{\prime}$ and $\phi_{m+1}^{\prime}, \ldots, \phi_{p}^{\prime}$ be constructed as above; then

$$
\phi(z)^{\prime}=\left(\begin{array}{c}
\phi_{1}(z)^{\prime}  \tag{11}\\
\vdots \\
\phi_{m}(z)^{\prime}
\end{array}\right) \quad \text { and } \quad\binom{\phi(z)^{\prime}}{\hdashline a_{\perp}^{\prime}}=\left(\begin{array}{c}
\phi_{1}(z)^{\prime} \\
\vdots \\
\phi_{m}(z)^{\prime} \\
\hdashline \phi_{m+1}^{\prime} \\
\vdots \\
\phi_{p}^{\prime}
\end{array}\right)
$$

are called a canonical system of root functions (respectively an extended canonical system of root functions) of $C(z)$ at $z_{\omega}$ of orders $\left(s_{1}, s_{2}, \ldots, s_{m}\right)$ (respectively $\left(s_{1}, s_{2}, \ldots, s_{m}, s_{m+1}, \ldots, s_{p}\right)$ ) with $\infty>s_{1} \geq s_{2} \geq \cdots \geq s_{m}>0=s_{m+1}=\cdots=s_{p}$.

Such a canonical system of root functions is not unique. To see this, one can show that the first row vector $\phi_{1}(z)^{\prime}$ in (11) can be replaced by a combination of $\phi_{1}(z)^{\prime}$ and $\phi_{2}(z)^{\prime}$, called $\phi_{1}^{\star}(z)^{\prime}$, and the canonical system of root functions containing $\phi_{1}^{\star}(z)^{\prime}$ would still satisfy the definition. More specifically, define $\phi_{1}^{\star}(z)^{\prime}:=\phi_{1}(z)^{\prime}+\left(z-z_{\omega}\right)^{s_{1}-s_{2}} \phi_{2}(z)^{\prime}$ and observe that, by Definition $3, \phi_{j}(z)^{\prime} C(z)=\left(z-z_{\omega}\right)^{s_{j}} \widetilde{\phi}_{j}(z)^{\prime}$, with $\widetilde{\phi}_{j}\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}, j=1,2$. Hence one has

$$
\phi_{1}^{\star}(z)^{\prime} C(z)=\left(z-z_{\omega}\right)^{s_{1}} \widetilde{\phi}_{1}^{\prime}(z)+\left(z-z_{\omega}\right)^{s_{1}-s_{2}+s_{2}} \widetilde{\phi}_{2}^{\prime}(z)=\left(z-z_{\omega}\right)^{s_{1}} \widetilde{\phi}^{\star \prime}(z)
$$

where $\widetilde{\phi}^{\prime \star}(z):=\widetilde{\phi}_{1}^{\prime}(z)+\widetilde{\phi}_{2}^{\prime}(z)$. Because $\widetilde{\phi}_{j}\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}, j=1,2$, one has $\widetilde{\phi}^{\prime \star}\left(z_{\omega}\right) \neq 0^{\prime}$ unless $\widetilde{\phi}_{1}\left(z_{\omega}\right)^{\prime}=-\widetilde{\phi}_{2}\left(z_{\omega}\right)^{\prime}$. However, this last case is ruled out because it would contradict the fact that $s_{1}$ is maximal. Hence $\widetilde{\phi}^{\prime \star}\left(z_{\omega}\right) \neq 0^{\prime}$. This shows that $\phi_{1}^{\prime \star}(z)$ satisfies the definition of root function of order $s_{1}$, and hence it can replace $\phi_{1}^{\prime}(z)$ in (11).

While a canonical system of root functions (and also an extended canonical system of root functions) is not unique, the orders $s_{1} \geq s_{2} \geq \cdots \geq s_{m}>0=s_{m+1}=\cdots=s_{p}$ are uniquely determined by $C(z)$ at $z_{\omega}$, see Lemma 1.1 in Gohberg et al. (1993); they are called the partial multiplicities of $C(z)$ at $z_{\omega}$.

Finally, consider the local Smith factorization of $C(z)$ at $z=z_{\omega}$, see Gohberg et al. (1993), i.e., the factorization

$$
\begin{equation*}
C(z)=E(z) M(z) H(z), \tag{12}
\end{equation*}
$$

where $M(z)=\operatorname{diag}\left(\left(z-z_{\omega}\right)^{s_{h}}\right)_{h=1, \ldots, p}$ is uniquely defined and contains the partial multiplicities $s_{1} \geq \cdots \geq s_{p}$ of $C(z)$ at $z=z_{\omega}$; the matrices $E(z), H(z)$ are analytic and invertible
in a neighbourhood of $z=z_{\omega}$ and are non-unique. $M(z)$ is called the local Smith form of $C(z)$ at $z=z_{\omega} \cdot{ }^{6}$

Remark 22 (Extended canonical system of root functions in the I(1) VAR case). In the $I(1)$ VAR case, see Example 1, the orders of an extended canonical system of root functions of $C(z)$ at 1 are $\left(s_{1}, \ldots, s_{r_{0}}, s_{r_{0}+1}, \ldots, s_{p}\right)=(1, \ldots, 1,0, \ldots, 0)$ and a possible choice of an extended canonical system of root functions corresponding to these unique orders is given by the $p$ rows in $\left(\beta_{0}, \beta_{1}\right)^{\prime}$.

Remark 23 (Extended canonical system of root functions in the I(2) VAR case). In the I(2) VAR case, see Example 2, the orders of an extended canonical system of root functions of $C(z)$ at 0 are $\left(s_{1}, \ldots, s_{r_{0}}, s_{r_{0}+1}, \ldots, s_{r_{0}+r_{1}}, s_{r_{0}+r_{1}+1}, \ldots, s_{p}\right)=(2, \ldots, 2,1, \ldots, 1,0, \ldots, 0)$ and a possible choice of an extended canonical system of root functions corresponding to these unique orders is given by the $p$ rows in $\left(\beta_{0}+(1-z)\left(\bar{\alpha}_{0}^{\prime} A_{1}\right)^{\prime}, \beta_{1}, \beta_{2}\right)^{\prime}$.

## 5. Cointegrating Spaces

Let $\phi(z)^{\prime}$ be a canonical system of root functions of $C(z)$ at $z_{\omega}$, see Definition 4 . Appendix A. 2 shows that $\operatorname{row}_{G}\left(\phi(z)^{\prime}\right)$ with $G=F, F[z], F(z)$ are well defined sets of (generalized) root functions. This section argues that one could take any of them as a definition of 'cointegrating space' for multicointegrated systems. Note that

$$
\operatorname{row}_{F}\left(\phi(z)^{\prime}\right) \subset \operatorname{row}_{F[z]}\left(\phi(z)^{\prime}\right) \subset \operatorname{row}_{F(z)}\left(\phi(z)^{\prime}\right)
$$

so that the three definitions of cointegrating space are naturally nested. Remark that $\operatorname{row}_{F}\left(\phi(z)^{\prime}\right)$ is a vector space over $F, \operatorname{row}_{F[z]}\left(\phi(z)^{\prime}\right)$ is a free module over the ring $F[z]$ of polynomials in $z$ (which contains $\operatorname{row}_{F}\left(\phi(z)^{\prime}\right)$ ) and $\operatorname{row}_{F(z)}\left(\phi(z)^{\prime}\right)$ is a vector space over the field $F(z)$ of rationals functions of $z$ ((which contains $\operatorname{row}_{F[z]}\left(\phi(z)^{\prime}\right)$ and hence $\left.\operatorname{row}_{F}\left(\phi(z)^{\prime}\right)\right)$. Finally note the central role played by the canonical system of root functions $\phi(z)^{\prime}$ as a basis for these different spaces, which differ for the set of scalars chosen in linear combinations.

### 5.1. The Cointegrating Space $\operatorname{row}_{F}\left(\phi(z)^{\prime}\right)$ as a Vector Space over $F$

The cointegrating space $\operatorname{row}_{F}\left(\phi(z)^{\prime}\right)$, where $F=\mathbb{R}, \mathbb{C}$, is a vector space. In fact, the set of all $F$-linear combination of $\phi(z)^{\prime}$ produces a vector space, because $\operatorname{row}_{F}\left(\phi(z)^{\prime}\right)$ is closed under multiplication by a scalar in $F$ by Proposition A1 and with respect to vector addition, as a special case of Proposition A2.

In order to discuss the cointegrating spaces $\operatorname{row}_{F[z]}\left(\phi(z)^{\prime}\right)$ and $\operatorname{row}_{F(z)}\left(\phi(z)^{\prime}\right)$, the notion of generalized cointegrating vector is introduced, as the counterpart of the notion of generalized root function, see Definition A1.

Definition 5 (Generalized cointegrating vector at frequency $\omega$ ). Let $n \in \mathbb{Z}$ and $\zeta(z)^{\prime}$ be a cointegrating vector at frequency $\omega$ and order s, see Definition 2; then

$$
\xi(z)^{\prime}:=\left(1-z / z_{\omega}\right)^{n} \zeta(z)^{\prime}
$$

is called a generalized cointegrating vector at frequency $\omega$ with order s and exponent $n$.

### 5.2. The Cointegrating Space $\operatorname{row}_{F[z]}\left(\phi(z)^{\prime}\right)$ as a Free Module over $F[z]$

Consider next $\operatorname{row}_{F[z]}\left(\phi(z)^{\prime}\right) . F[z]$ is the polynomial ring formed as the set of polynomials in $z$ with coefficients in $F$. As it is well known, $F[z]$ is a ring but not a field (division ring), see e.g., Hungerford (1980), because polynomials, unlike rational functions, lack the multiplication inverse. The following propositions summarizes that $\operatorname{row}_{F[z]}\left(\phi(z)^{\prime}\right)$ is a free module over the ring $F[z]$ of polynomials.

Proposition 3 (row $\operatorname{ron}_{F[z]}\left(\phi(z)^{\prime}\right)$ is a $F[z]$-module). Consider $\mathcal{G}=\operatorname{row}_{F[z]}\left(\phi(z)^{\prime}\right)$, where $\phi(z)^{\prime}$ is a canonical system of root functions of $C(z)$ at $z_{\omega}$ with coefficients in $F$, and where $F[z]$ is the
ring of polynomials in $z$ with coefficients in $F$; then $\mathcal{G}$ is closed with respect to the vector sum, and it is closed under multiplication by a scalar polynomial in $F[z]$; hence $\mathcal{G}$ is a module over the ring $F[z]$ of polynomials.

Proof. By Propositions A1 and A2, $\mathcal{G}$ is closed under addition and under multiplication by a scalar polynomial in $F[z]$. One needs to verify that, see e.g., Definition IV.1.1 in Hungerford (1980), for $\zeta(z), \psi(z) \in \mathcal{G}$ and $1, a(z), b(z) \in F[z]$

$$
\begin{align*}
a(z) \cdot\left(\zeta(z)^{\prime}+\psi(z)^{\prime}\right) & =a(z) \cdot \zeta(z)^{\prime}+a(z) \cdot \psi(z)^{\prime} \\
(a(z)+b(z)) \cdot \zeta(z)^{\prime} & =a(z) \cdot \zeta(z)^{\prime}+b(z) \cdot \zeta(z)^{\prime}  \tag{13}\\
(a(z) b(z)) \cdot \zeta(z)^{\prime} & =a(z) \cdot\left(b(z) \cdot \zeta(z)^{\prime}\right) \\
1 \cdot \zeta(z)^{\prime} & =\zeta(z)^{\prime}
\end{align*}
$$

where • indicates multiplication by a scalar. The distributive properties in (13) are seen to be satisfied. This proves the statement.

### 5.3. The Cointegrating Space $\operatorname{row}_{F(z)}\left(\phi(z)^{\prime}\right)$ as a Vector Space over $F(z)$

Finally consider $\operatorname{row}_{F(z)}\left(\phi(z)^{\prime}\right)$. The set of scalars $F(z)$ is the field of rational functions in $z$ with coefficients in $F$. As it is well known, $F(z)$ is a field (division ring), see e.g., Hungerford (1980).

Remark 24 (Rational vectors without poles at $z_{\omega}$ ). Take $\zeta(L)^{\prime}$ to be a rational vector, i.e., of the form $\zeta(z)^{\prime}=\frac{1}{d(z)} b(z)^{\prime}$ where $d(z)$ is a monic polynomial and $b(z)^{\prime}$ is a $1 \times p$ vector polynomial, with $d(z)$ and $b(z)^{\prime}$ relatively prime, see Example A1. If $d(z)$ has no root equal to $z_{\omega}$, then $\zeta(z)^{\prime}$ is an analytic function on $D\left(z_{\omega}, \eta\right), \eta>0$, see Remark A1 and Lemma A1; hence a special case of an analytic vector function $\zeta(z)^{\prime}$ is a rational vector with denominator $d(z)$ without roots equal to $z_{\omega}$.

Remark 25 (Rational vectors with poles at $z_{\omega}$ ). If $d(z)$ has one root equal to $z_{\omega}$ with multiplicity $m$, then $\zeta(z)^{\prime}$ has a pole of order $m$, and it is not an analytic function on some $D\left(z_{\omega}, \eta\right), \eta>0$; hence Definition 2 cannot be applied, because it requires $\zeta(z)^{\prime}$ to be analytic. However, one could remove the pole of order $m$ by defining $\xi(z)^{\prime}:=\left(1-z / z_{\omega}\right)^{m} \zeta(z)^{\prime}$, and use Definition 2 on $\xi(z)^{\prime}$, which is analytic function, as done in Definition 5.

Remark 26 (Representation for generic rational vectors). In the following, when dealing with rational vectors of the type $\zeta(z)^{\prime}=\frac{1}{d(z)} b(z)^{\prime}$, it is sufficient to consider the case where $d(z)$ does not have a root at $z_{\omega}$, thanks to Definition 5. In fact, let $d(z)$ be decomposed as $d(z)=$ $\left(1-z / z_{\omega}\right)^{m} d_{\star}(z)$ with $d_{\star}\left(z_{\omega}\right) \neq 0$ and $m \geq 0$; in this representation, $z_{\omega}$ is a root of $d(z)$ if and only if $m>0$ and it is not a root if and only if $m=0$. By Remark $24, \zeta(z)^{\prime}$ is a (generalized) cointegrating vector if and only if $\xi(z)^{\prime}:=\left(1-z / z_{\omega}\right)^{m} \zeta(z)^{\prime}=\frac{1}{d_{\star}(z)} b(z)^{\prime}$ is a cointegrating vector. Hence Definition 5 allows to concentrate on the case where the denominator has no root at $z_{\omega}$.

The following proposition summarizes that $\operatorname{row}_{F(z)}\left(\phi(z)^{\prime}\right)$ is a vector space over the field $F(z)$ of rational functions.

Proposition $4\left(\operatorname{row}_{F(z)}\left(\phi(z)^{\prime}\right)\right.$ is a vector space over $F(z)$ ). Let $\mathcal{H}=\operatorname{row}_{F(z)}\left(\phi(z)^{\prime}\right)$ where $\phi(z)^{\prime}$ is a canonical system of root functions of $C(z)$ at $z_{\omega}$ with coefficients in $F$, where $F(z)$ is the field of rational function in $z$ with coefficients in $F$; then $\mathcal{H}$ is closed with respect to the vector sum, and under multiplication by a scalar rational function in $F(z)$, and $\mathcal{H}$ is a vectors space over the field $F(z)$ of rational functions.

Proof. $\mathcal{H}$ is closed with respect to multiplication by a rational function in $F(z)$, see Proposition A1, and with respect to vector addition, see Proposition A2. One can ver-
ify for $\zeta(z), \psi(z) \in \mathcal{H}$ and $1, a(z), b(z) \in F(z)$, that the distribution equalities in (13) are satisfied. Because $F(z)$ is a field, then $\mathcal{H}$ is a vector space over $F(z)$.

## 6. The Local Rank Factorization

This section shows how to explicitly obtain a canonical system of root functions $\phi(z)^{\prime}$ or an extended canonical system of root functions $\left(\phi(z), a_{\perp}\right)^{\prime}$ for a generic VAR process

$$
\begin{equation*}
A(L) x_{t}=\varepsilon_{t}, \quad A_{0} \neq 0, \quad \operatorname{det} A_{0}=0 \tag{14}
\end{equation*}
$$

with $A(z)$ analytic for all $z \in D(0,1+\eta), \eta>0$, having roots at $z=z_{\omega}=e^{i \omega}$ and at $z$ with $|z|>1$, see Remarks 1 and 2 .

The derivation of the Granger representation theorem involves the inversion of the matrix function

$$
\begin{equation*}
A(z)=\sum_{n=0}^{\infty}\left(z-z_{\omega}\right)^{n} A_{n}, \quad A_{n} \in \mathbb{C}^{p \times p}, \quad A_{0} \neq 0, \quad \operatorname{det} A_{0}=0 \tag{15}
\end{equation*}
$$

in $D\left(z_{\omega}, \eta\right)$. This includes the case of matrix polynomials $A(z)$, in which the degree of $A(z)$ is finite, $k$ say, with $A_{n}=0$ for $n>k .{ }^{7}$

The inversion of $A(z)$ around the singular point $z=z_{\omega}$ yields an inverse with a pole of some order $d=1,2, \ldots$ at $z=z_{\omega}$; an explicit condition on the coefficients $\left\{A_{n}\right\}_{n=0}^{\infty}$ in (15) for $A(z)^{-1}$ to have a pole of given order $d$ is described in Theorem 2 below; this is indicated as the $\operatorname{POLE}(d)$ condition in the following. Under the POLE $(d)$ condition, $A(z)^{-1}$ has Laurent expansion around $z=z_{\omega}$ given by

$$
\begin{equation*}
A(z)^{-1}=:\left(z-z_{\omega}\right)^{-d} C(z)=\sum_{n=0}^{\infty}\left(z-z_{\omega}\right)^{n-d} C_{n}, \quad C_{0} \neq 0, \quad \operatorname{det} C_{0}=0 \tag{16}
\end{equation*}
$$

Note that $C\left(z_{\omega}\right)=C_{0} \neq 0$ and $C(z)$ is expanded around $z=z_{\omega}$. In the following, the coefficients $\left\{C_{n}\right\}_{n=0}^{\infty}$ are called the Laurent coefficients. The first $d$ of them, $\left\{C_{n}\right\}_{n=0}^{d-1}$, make up the principal part and characterize the singularity of $A(z)^{-1}$ at $z=z_{\omega}$.

The following result is taken from Franchi and Paruolo (2019) Theorem 3.3. ${ }^{8}$
Theorem 2 (POLE $(d)$ condition). Consider $A(z)$ defined in (15); let $0<r_{0}:=\operatorname{rank} A_{0}<p$, $r_{0}^{\max }:=p$ and define $\alpha_{0}, \beta_{0}$ by the rank factorization $A_{0}=-\alpha_{0} \beta_{0}^{\prime}$. Moreover, for $j=1,2, \ldots$ define $\alpha_{j}, \beta_{j}$ by the rank factorization

$$
\begin{equation*}
P_{a_{j \perp}} A_{j, 1} P_{b_{j \perp}}=-\alpha_{j} \beta_{j}^{\prime}, \quad a_{j}:=\left(\alpha_{0}, \ldots, \alpha_{j-1}\right), \quad b_{j}:=\left(\beta_{0}, \ldots, \beta_{j-1}\right) \tag{17}
\end{equation*}
$$

where $P_{x}$ denotes the orthogonal projection onto the space spanned by the columns of $x$ and

$$
A_{h+1, n}:=\left\{\begin{array}{cl}
A_{n} & \text { for } h=0  \tag{18}\\
A_{h, n+1}+A_{h, 1} \sum_{i=0}^{h-1} \bar{\beta}_{i} \bar{\alpha}_{i}^{\prime} A_{i+1, n} & \text { for } h=1,2, \ldots
\end{array}, \quad n=0,1, \ldots\right.
$$

Finally, let

$$
\begin{equation*}
r_{j}:=\operatorname{rank}\left(P_{a_{j \perp}} A_{j, 1} P_{b_{j \perp}}\right), \quad r_{j}^{\max }:=p-\sum_{i=0}^{j-1} r_{i} . \tag{19}
\end{equation*}
$$

Then, a necessary and sufficient condition for $A(z)$ to have an inverse with pole of order $d=1,2, \ldots$ at $z=z_{\omega}$-called $\operatorname{POLE}(d)$ condition - is that

$$
\left\{\begin{array}{ll}
r_{j}<r_{j}^{\max } & \text { (reduced rank condition) for } j=0, \ldots, d-1 \\
r_{d}=r_{d}^{\max } & \text { (full rank condition) for } j=d
\end{array} .\right.
$$

Observe that because rank $P_{a_{j \perp}} A_{j, 1} P_{b_{j \perp}}=\operatorname{rank} a_{j \perp}^{\prime} A_{j, 1} b_{j \perp}$, one has $r_{j}=\operatorname{rank} a_{j \perp}^{\prime} A_{j, 1} b_{j \perp} ;$ hence $d=1$ if and only if

$$
r_{1}=r_{1}^{\max }, \quad \text { where } \quad r_{1}=\operatorname{rank} \alpha_{0 \perp}^{\prime} A_{1} \beta_{0 \perp} \quad \text { and } \quad r_{1}^{\max }=p-r_{0}
$$

This corresponds to the condition in Howlett (1982, Theorem 3) and to the $I(1)$ condition in Johansen (1991, Theorem 4.1). Similarly, one has $d=2$ if and only if $r_{1}<r_{1}^{\max }$,

$$
r_{2}=r_{2}^{\max }, \quad \text { where } \quad r_{2}=\operatorname{rank} a_{2 \perp}^{\prime}\left(A_{2}+A_{1} \bar{\beta}_{0} \bar{\alpha}_{0}^{\prime} A_{1}\right) b_{2 \perp} \quad \text { and } \quad r_{2}^{\max }=p-r_{0}-r_{1}
$$

which corresponds to the $I(2)$ condition in Johansen (1992, Theorem 3).
Theorem 2 is thus a generalization of the Johansen's $I(1)$ and $I(2)$ conditions and shows that, in order to have a pole of order $d$ in the inverse, one needs $d+1$ rank conditions on $A(z)$ : The first $j=0, \ldots, d-1$ are reduced rank conditions, $r_{j}<r_{j}^{\max }$, that establish that the order of the pole is greater than $j$; the last one is a full rank condition, $r_{d}=r_{d}^{\max }$, that establishes that the order of the pole is exactly equal to $d$. These requirements make up the POLE $(d)$ condition.

The following result is also taken from Franchi and Paruolo (2019). ${ }^{9}$
Theorem 3 (Local Smith factorization). Consider $A(z)$ and the other related quantities defined in Theorem 16; for $j=0, \ldots, d$, define the $r_{j} \times p$ matrix functions $\gamma_{j}(z)^{\prime}$ as follows

$$
\begin{equation*}
\gamma_{j, 0}^{\prime}:=\beta_{j}^{\prime}, \quad \gamma_{j, n}^{\prime}:=-\bar{\alpha}_{j}^{\prime} A_{j+1, n}, \quad n=1,2, \ldots \quad \gamma_{j}(z)^{\prime}:=\sum_{n=0}^{\infty}\left(z-z_{\omega}\right)^{n} \gamma_{j, n}^{\prime} \tag{20}
\end{equation*}
$$

and define the $p \times p$ matrix functions $\Gamma(z)$ and $\Lambda(z)$ as follows

$$
\Gamma(z):=\left(\begin{array}{c}
\gamma_{0}(z)^{\prime}  \tag{21}\\
\vdots \\
\gamma_{d}(z)^{\prime}
\end{array}\right), \quad \Lambda(z):=\left(\begin{array}{ccc}
\left(z-z_{\omega}\right)^{0} I_{r_{0}} & & \\
& \ddots & \\
& & \left(z-z_{\omega}\right)^{d} I_{r_{d}}
\end{array}\right)
$$

Then $\Gamma(z), \Xi(z):=A(z) \Gamma(z)^{-1} \Lambda(z)^{-1}$ are analytic and invertible on $D\left(z_{\omega}, \eta\right), \eta>0$, and $\Lambda(z)$ is the local Smith form of $A(z)$ at $z_{\omega}, A(z)=\Xi(z) \Lambda(z) \Gamma(z)$. Moreover one can choose the factors $E(z), M(z), H(z)$ for the local Smith factorization of $C(z)$ defined in (16), see (12), as

$$
E(z)=\Gamma(z)^{-1}, \quad M(z)=\left(z-z_{\omega}\right)^{d} \Lambda(z)^{-1}, \quad H(z)=\Xi(z)^{-1}
$$

Theorem 3 shows that the LRF fully characterizes the elements of the local Smith factorization of $C(z)$ at $z_{\omega}$. In fact, the values of $j$ with $r_{j}>0$ in the LRF provide the distinct partial multiplicities of $C(z)$ at $z_{\omega}$ and $r_{j}$ gives the number of partial multiplicities that are equal to a given $j$; this characterizes the local Smith form $\Lambda(z)$. Moreover, it also provides the constructions of an extended canonical system of root functions.

Remark that the $j$-th block of rows in $\Gamma(z) C(z)=\left(z-z_{\omega}\right)^{d} \Lambda(z)^{-1} \Xi(z)^{-1}$ can be written as

$$
\begin{equation*}
\gamma_{j}(z)^{\prime} C(z)=\left(z-z_{\omega}\right)^{d-j} \widetilde{\gamma}_{j}(z)^{\prime}, \quad j=0, \ldots, d \tag{22}
\end{equation*}
$$

where $\gamma_{j}\left(z_{\omega}\right)^{\prime}=\beta_{j}^{\prime}$ and $\widetilde{\gamma}_{j}\left(z_{\omega}\right)^{\prime}$ have full row rank; here $\widetilde{\gamma}_{j}(z)^{\prime}$ denotes the corresponding block of rows in $\Xi(z)^{-1}$. This shows that $\gamma_{j}(z)^{\prime}$ are $r_{j}$ root functions of order $d-j$ of $C(z)$.

The next result presents the Triangular representation as proved in Franchi and Paruolo (2019, Corollary 4.6).

Proposition 5 (Triangular representation). Let $x_{t}$ in (14) satisfy the POLE $(d)$ condition on $A(z)$ and define

$$
\begin{align*}
\Gamma_{\circ}(L) & :=\binom{\phi(L)^{\prime}}{\left.\hdashline-\bar{\beta}_{d}^{\prime}--\right)} \\
\phi(L)^{\prime} & :=\left(\begin{array}{c}
\gamma_{0}^{(d-1)}(L)^{\prime} \\
\gamma_{1}^{(d-2)}(L)^{\prime} \\
\vdots \\
\gamma_{d-1}^{(0)}(L)^{\prime}
\end{array}\right)=\left(\begin{array}{c}
\beta_{0}^{\prime}+\sum_{k=1}^{d-1}\left(-z_{\omega}\right)^{k} \gamma_{0, k}^{\prime} \Delta_{\omega}^{k} \\
\beta_{1}^{\prime}+\sum_{k=1}^{d-2}\left(-z_{\omega}\right)^{k} \gamma_{1, k}^{\prime} \Delta_{\omega}^{k} \\
\vdots \\
\beta_{d-1}^{\prime}
\end{array}\right), \tag{23}
\end{align*}
$$

where $\gamma_{j}^{(d-j-1)}(z)^{\prime}=\sum_{k=0}^{d-j-1} \gamma_{j, k}^{\prime}\left(z-z_{\omega}\right)^{k}$ is the truncation of order $d-j-1$ of the root functions $\gamma_{j}(z)^{\prime}$ in (20). Then $x_{t}$ is $I(d)$ and it admits the Triangular Representation

$$
\Lambda(L) \Gamma_{\circ}(L) x_{t} \sim I(0)
$$

where no linear combination exists of the l.h.s. that is integrated of lower order.
Observe that the canonical system of root functions $\phi(z)^{\prime}$ in (23) is not unique and not of minimal polynomial order, as discussed in the next section. The following example applies the above concepts in the $I(2)$ VAR case.

Example 3 (I(2) VAR example continued). Consider Example 2. Applying truncation to the rows of $\left(\beta_{0}^{\prime}+\Delta \bar{\alpha}_{0}^{\prime} A_{1}\right)$, see Propositions 5 and $A 3$, one finds that the columns in $\beta_{0}^{\prime}$ are root functions of $C(z)$ at $\omega=0$ of order at least $\min (2,1)=1$. Consider now one row in $\left(\beta_{0}^{\prime}+\Delta \bar{\alpha}_{0}^{\prime} A_{1}+\Delta^{2} A^{\prime}\right)$ for some matrix $A$; this root function is of order 2 by Remark 23, and its truncation to degree 1, i.e., to the corresponding row of $\left(\beta_{0}^{\prime}+\Delta \bar{\alpha}_{0}^{\prime} A_{1}\right)$ is still of order 2 by Propositions 5 and A3, Finally consider one row in $\left(\beta_{0}^{\prime}+\Delta A^{\prime}\right)$, which gives a root function of order at least 1 ; its truncation to a polynomial of degree 0 gives the corresponding row of $\beta_{0}^{\prime}$, which has order at least 1 by Propositions 5 and $A 3$. In fact the rows of $\beta_{0}^{\prime}$ give root functions of order equal to 1 or to 2 , when the corresponding entries in $\bar{\alpha}_{0}^{\prime} A_{1}$ in $\left(\beta_{0}^{\prime}+\Delta \bar{\alpha}_{0}^{\prime} A_{1}\right)$ are equal to 0 , as discussed below.

## 7. Minimal Bases

This section describes the algorithm of Forney (1975) to reduce the basis $\phi(z)^{\prime}$ to minimal order, using the generic notation of $b(z)^{\prime}$ in place of $\phi(z)^{\prime}$. The generic basis $b(z)^{\prime}$ is assumed to be rational and of dimension $r \times p$. This algorithm exploits the nesting $\operatorname{row}_{F}\left(b(z)^{\prime}\right) \subset \operatorname{row}_{F[z]}\left(b(z)^{\prime}\right) \subset \operatorname{row}_{F(z)}\left(b(z)^{\prime}\right)$. In the following, the $j$-row of $b(z)^{\prime}$ is indicated as $b_{j}(z)^{\prime}$, which is the $j$-th element of the basis, $j=1, \ldots, r$. Various modifications of the original basis $b^{(0)}(z)^{\prime}:=b(z)^{\prime}$ are indicated as $b^{(h)}(z)^{\prime}$ for $h=1,2,3$.

Definition 6 (Degree of $\left.b(z)^{\prime}\right)$. If $b(z)^{\prime}$ is a polynomial basis, the degree $v_{j}$ of its $j$-th row, indicated as $v_{j}:=\operatorname{deg} b_{j}(z)^{\prime}$, is defined as the maximum degree of its elements, and the degree $v$ of $b(z)^{\prime}$ is defined as $v:=\operatorname{deg} b(z)^{\prime}:=\sum_{j=1}^{r} v_{j}$, i.e., the sum of the degrees of its rows.

The reduction algorithm proposed by Forney (1975, pp. 497-98) consists of the following 3 steps.
Step 1 If $b^{(0)}(z)^{\prime}$ is not polynomial, multiply each row by the least common denominator of each row to obtain a polynomial basis $b^{(1)}(z)^{\prime}$.
Step 2 Reduce row orders in $b^{(1)}(z)^{\prime}$ by taking $F[z]$-linear combinations.
Step 3 Reduce $b^{(2)}(z)^{\prime}$ to a basis $b^{(3)}(z)^{\prime}$ with a full-row-rank high order coefficient matrix, i.e., a "row proper" basis.

This procedure gives a final basis $b^{(3)}(z)^{\prime}$ which has lowest degree, see Forney (1975) Section 3.

Remark 27 (Spaces and algorithm). Step 1 works on $\operatorname{row}_{F(z)}\left(\phi(z)^{\prime}\right)$, Step 2 works on $\operatorname{row}_{F[z]}\left(\phi(z)^{\prime}\right)$, Step 3 uses F-linear combinations on $Q(z) \phi(z)^{\prime}$ with appropriate square polynomial matrices $Q(z)$.

### 7.1. Step 1

If $b^{(0)}(z)^{\prime}$ is polynomial, the algorithm sets $b^{(1)}(z)^{\prime}=b^{(0)}(z)^{\prime}$; otherwise $b^{(0)}(z)^{\prime}$ is rational, and its $j$-th row $b_{j}(z)^{\prime}$ has representation $b_{j}(z)^{\prime}=\frac{1}{a_{j}(z)} c_{j}(z)^{\prime}$, where $c_{j}(z)^{\prime}$ is a polynomial row vector and $a_{j}(z)$ is a scalar polynomial, and $c_{j}(z)$ and $a_{j}(z)$ are relatively prime. The first step consist in computing $b^{(1)}(z)^{\prime}=\operatorname{diag}\left(a_{1}(z), \ldots, a_{r}(z)\right) b^{(0)}(z)^{\prime}$, where $Q(z):=\operatorname{diag}\left(a_{1}(z), \ldots, a_{r}(z)\right)$ is a square polynomial matrix of dimension $r$.

### 7.2. Step 2

The second step reduces the degree of the rows in $b^{(1)}(z)^{\prime}$. This involves finding specific points $z_{h}, h=1, \ldots, k$, at which $\operatorname{rank}\left(b^{(1)}\left(z_{h}\right)^{\prime}\right)<r$. To find them, one can calculate the greatest common divisor $\ell(z)$ of all $r \times r$ minors of $b^{(1)}(z)^{\prime}$. If $\ell(z)=1$ this step is complete, and the algorithm sets $b^{(2)}(z)^{\prime}=b^{(1)}(z)^{\prime}$; otherwise one computes the zeros of $\ell(z), z_{1}, \ldots, z_{k}$ say, where $z_{h} \in \mathbb{C}, h=1, \ldots, k$. The following substep is then applied to each root $z_{h}$ sequentially, $h=1, \ldots, k$.

Denote by $w(z)^{\prime}$ the current basis; this will be replaced by $\kappa(z)^{\prime}$ at the end of this substep. For $h=1$, one has $w(z)^{\prime}=b^{(1)}(z)^{\prime}$. For $z=z_{h}$, all minors of order $r$ of $w\left(z_{h}\right)^{\prime}$ vanish, which means that $w\left(z_{h}\right)^{\prime}$ is singular, i.e., it has reduced rank and rank factorization $w\left(z_{h}\right)^{\prime}=\psi a^{\prime}$, say, where $\psi, a$ are full column rank. Let $c^{\prime}:=\left(c_{1}, \ldots, c_{p}\right)$ be one row in $\psi_{\perp}^{\prime}$. Indicate by $A_{c}:=\left\{i: c_{i} \neq 0\right\}$ the set of its non-zero coefficients, and let $v_{i_{0}}:=\max _{i \in A_{c}}\left\{v_{i}\right\}$ be the maximal degree of rows in $w(z)^{\prime}$ with nonzero coefficient in $c^{\prime}$.

This substep consists of replacing row $i_{0}$ of $w(z)^{\prime}$ with $c^{\prime} w(z)^{\prime} /\left(z-z_{h}\right)$, which is still a polynomial vector. In fact $c^{\prime} w\left(z_{h}\right)^{\prime}=0^{\prime}$, so that $c^{\prime} w\left(z_{h}\right)^{\prime}$ has representation $c^{\prime} w\left(z_{h}\right)^{\prime}=$ $\left(z-z_{h}\right) \tau(z)^{\prime}$ with $\tau(z)^{\prime}$ a polynomial vector, so that $c^{\prime} w(z)^{\prime} /\left(z-z_{h}\right)=\tau(z)^{\prime}$. This defines $\kappa(z)^{\prime}$ in terms of $w(z)^{\prime}$ as $\kappa(z)^{\prime}=B(z)^{-1} Q w(z)^{\prime}$ where $Q$ is an $r \times r$ square matrix, equal to $I_{r}$ except for row $i_{0}$, equal to $c^{\prime}$, and where $B(z)$ is a diagonal matrix equal to $I_{r}$ except for having $z-z_{h}$ in its $i_{0}$-th position on the diagonal. Note that $Q$ is nonsingular, because $c_{i_{0}} \neq 0$. The same procedure is applied to each row $c^{\prime}$ of $\psi_{\perp}^{\prime}$.

This substep is repeated for all $z_{j}, j=1, \ldots, k$. The condition on the minors in then recalculated and the substep repeated for the new roots, until the greatest common divisor $\ell(z)$ of all $r \times r$ minors of $\kappa(z)^{\prime}$ is 1 . When this is the case, Step 2 sets $b^{(2)}(z)^{\prime}=\kappa(z)^{\prime}$.

### 7.3. Step 3

The last step operates on the high order coefficient matrix, repeating the following substep. Let $w(z)^{\prime}$ indicate $b^{(2)}(z)^{\prime}$ at the beginning of the substep, which will be replaced by $\kappa(z)^{\prime}$ at the end of it. Let $v_{i}$ be the order of the $i$-th row of $w(z)^{\prime}$, indicated as $w_{i}(z)^{\prime}=\sum_{j=0}^{v_{i}}\left(z-z_{\omega}\right)^{j} w_{i j}^{\prime}$. The high-order matrix is defined as the $r \times p$ matrix $w_{*}^{\prime}:=\left(w_{1 v_{1}}, \ldots, w_{r v_{r}}\right)^{\prime}$ composed of the coefficient matrix of the highest degree of $\left(z-z_{\omega}\right)$ for each row of $w(z)^{\prime}$.

A necessary and sufficient condition for $w_{*}^{\prime}$ to be of full rank is that the order of $w(z)^{\prime}$ is equal to the maximum order of its $r \times r$ minors. If this is not the case, $w_{*}^{\prime}$ is singular, i.e., it has rank factorization $w_{*}^{\prime}=\psi a^{\prime}$ with $\psi$ and $a$ of full column rank. Hence one can choose a vector $c^{\prime}:=\left(c_{1}, \ldots, c_{p}\right)$ as one row in $\psi_{\perp}^{\prime}$ for which one has $c^{\prime} w_{*}^{\prime}=0^{\prime}$.

As before, let $A_{c}:=\left\{i: c_{i} \neq 0\right\}$ and define $v_{i_{0}}:=\max _{i \in A_{c}}\left\{v_{i}\right\}$. Let also $n_{i}:=v_{i_{0}}-v_{i}$, note that $n_{i} \geq 0$ for $i \in A_{c}$ and let $Q(z):=\operatorname{diag}\left(\left(z-z_{\omega}\right)^{n_{1}}, \ldots\left(z-z_{\omega}\right)^{n_{r}}\right)$. Row $i_{0}$ in $w(z)^{\prime}$ is replaced by

$$
\begin{align*}
q(z)^{\prime}:=c^{\prime} Q(z) w(z)^{\prime} & =\sum_{i \in A_{c}} c_{i} \sum_{j=0}^{v_{i}}\left(z-z_{\omega}\right)^{j+n_{i}} w_{i, j}^{\prime}=\sum_{i \in A_{c}} c_{i} \sum_{s=n_{i}}^{v_{i_{0}}}\left(z-z_{\omega}\right)^{s} w_{i, s-n_{i}}^{\prime}  \tag{24}\\
& =\sum_{j=0}^{v_{i_{0}}-1}\left(z-z_{\omega}\right)^{j} q_{j}^{\prime}+\left(z-z_{\omega}\right)^{v_{i_{0}}} c^{\prime} w_{*}^{\prime}=\sum_{j=0}^{v_{i_{0}}-1}\left(z-z_{\omega}\right)^{j} q_{j}^{\prime}, \tag{25}
\end{align*}
$$

where $s$ in the last expression in the first line is defined as $j+n_{i}$ and $q_{j}^{\prime}:=\sum_{i \in A_{c}} c_{i} w_{i, n_{i}+j}^{\prime}$.
The central expression in (24) shows that $q(z)^{\prime}$ is polynomial because $n_{i} \geq 0$ in the exponents of $\left(z-z_{\omega}\right)$. In order to see that the degree of $q(z)^{\prime}$ is also lower than $v_{i_{0}}$, one can note that the the high order coefficient in (25), which correspond to $s=v_{i_{0}}$ in (24), equals $\sum_{i \in A_{c}} c_{i} w_{i, v_{i}}^{\prime}=c^{\prime} w_{*}^{\prime}=0^{\prime}$. This implies that the order of $q(z)^{\prime}$ is lower than $v_{i_{0}}$, and that replacing row $i_{0}$ of $w(z)^{\prime}$ with $q(z)^{\prime}$ reduces the order of the vector.

This defines $\kappa(z)^{\prime}$ in terms of $w(z)^{\prime}$ as $\kappa(z)^{\prime}=N Q(z) w(z)^{\prime}$ where $N$ is an $r \times r$ square matrix, equal to $I_{r}$ except for row $i_{0}$, equal to $c^{\prime}$. Note that $N$ is nonsingular, because $c_{i_{0}} \neq 0$. This process is repeated for all the rows $c^{\prime}$ in $\psi^{\prime}$. Next set $w(z)^{\prime}=\kappa(z)^{\prime}$ and repeat until the high order coefficient matrix has full rank. When this is the case, Step 3 sets $b^{(3)}(z)^{\prime}=\kappa(z)^{\prime}$.

## 8. From a Canonical System of Root Functions to a Minimal Basis for I(2) VAR

This section applies the algorithm of Forney reviewed in Section 7 to $\phi(z)^{\prime}$ in (23) to reduce the basis to minimal order in the $I(2)$ VAR example at frequency $\omega=0$. This application leads to the separation of the cases of
(i) non-polynomial cointegrating relations reducing the order of integration from 2 to 0 ;
(ii) polynomial cointegrating relations reducing the order of integration from 2 to 0 .

The process of obtaining minimal bases does not lead to a unique choice of basis; this leaves open the choice of how to further restrict the basis to obtain uniqueness. Forney (1975) obtains uniqueness requiring the minimal basis to be in upper echelon form. Other sets of restrictions can also be considered. For the sake of brevity, the restrictions on how to obtain a unique minimum basis are not further discussed here.

### 8.1. Step 1 in $I(2)$ VAR

Consider the triangular representation of an $I(2)$ system, see (23):

$$
\begin{equation*}
\Gamma_{\circ}(z):=\binom{\phi(z)^{\prime}}{\hdashline-\bar{\beta}_{2}^{\prime}}, \quad \phi(z)^{\prime}:=\binom{\gamma_{0}^{(1)}(z)^{\prime}}{\gamma_{1}^{(0)}(z)^{\prime}}=\binom{\beta_{0}^{\prime}+\gamma_{0,1}^{\prime}(z-1)}{\beta_{1}^{\prime}} \tag{26}
\end{equation*}
$$

and apply the algorithm of Forney (1975) to $b^{(0)}(z)^{\prime}:=\phi(z)^{\prime}$. Because $b^{(0)}(z)^{\prime}$ is already polynomial, one has $b^{(1)}(z)^{\prime}=b^{(0)}(z)^{\prime}=\phi(z)^{\prime}$.

### 8.2. Step 2 in I(2) VAR

Next consider Step 2, and set $w(z)^{\prime}=b^{(1)}(z)^{\prime}$. One wishes to find some zero $z_{h}$ and some corresponding $c^{\prime}$ so as have $c^{\prime} w\left(z_{h}\right)^{\prime}=0^{\prime}$. Denoting $u=z_{h}-1$, one hence needs to find the pair $\left(u, c^{\prime}\right)$ such that

$$
\begin{equation*}
c^{\prime}\left(\binom{\beta_{0}^{\prime}}{\beta_{1}^{\prime}}+\binom{\gamma_{0,1}^{\prime}}{0} u\right)=0 \tag{27}
\end{equation*}
$$

where $u$ is a scalar. Note that $u=0$ is not a possible zero of (27), because $w\left(z_{\omega}\right)=\left(\beta_{0}, \beta_{1}\right)^{\prime}$ is of full row rank, so that $u \neq 0$. Post-multiplying (27) by the square non-singular matrix $\left(\bar{\beta}_{0}, \bar{\beta}_{1}, \bar{\beta}_{2}\right)$ one finds

$$
c^{\prime}\left(\left(\begin{array}{ccc}
I_{r_{0}} & 0 & 0  \tag{28}\\
0 & I_{r_{1}} & 0
\end{array}\right)+\left(\begin{array}{ccc}
\gamma_{0,1}^{\prime} \bar{\beta}_{0} & \gamma_{0,1}^{\prime} \bar{\beta}_{1} & \gamma_{0,1}^{\prime} \bar{\beta}_{2} \\
0 & 0 & 0
\end{array}\right) u\right)=0 .
$$

Hence, partitioning $c^{\prime}$ as $c^{\prime}=\left(\varsigma^{\prime}, \theta^{\prime}\right)$ where $\varsigma^{\prime}$ is $1 \times r_{0}$, one finds that the second set of equations gives $\theta^{\prime}=0^{\prime}$ and the first one, substituting the expression of $\gamma_{0,1}^{\prime}=-\bar{\alpha}_{0}^{\prime} A_{1}$ given in Theorem 3, implies

$$
\begin{align*}
\varsigma^{\prime} \alpha_{0}^{\prime} A_{1} \bar{\beta}_{0} & =\lambda \varsigma^{\prime}, \quad \lambda:=u^{-1} \neq 0  \tag{29}\\
\varsigma^{\prime} \bar{\alpha}_{0}^{\prime} A_{1}\left(\bar{\beta}_{1}, \bar{\beta}_{2}\right) & =0, \tag{30}
\end{align*}
$$

where $\lambda=u^{-1} \neq 0$ in (29); note also that $u \neq 0$ has been simplified in (30). This proves the following proposition.

Proposition 6 (Step 2 condition in $I(2)$ ). A necessary and sufficient condition for Step 2 to be non-empty is that (29), (30) hold simultaneously, i.e., that $\left(\lambda, \varsigma^{\prime}\right)$ is a non-zero eigenvalue-left eigenvector pair of $\bar{\alpha}_{0}^{\prime} A_{1} \bar{\beta}_{0}$, and the left eigenvector $v^{\prime}$ is orthogonal to $\bar{\alpha}_{0}^{\prime} A_{1}\left(\bar{\beta}_{1}, \bar{\beta}_{2}\right)$. If this is the case, for each pair $\left(\lambda, \varsigma^{\prime}\right)$ one has

$$
\begin{equation*}
\varsigma^{\prime} \bar{\alpha}_{0}^{\prime} A_{1}=\varsigma^{\prime} \bar{\alpha}_{0}^{\prime} A_{1} P_{\beta_{0}}=\lambda \varsigma^{\prime} \beta_{0}^{\prime} . \tag{31}
\end{equation*}
$$

Observe that from (27), using $c^{\prime}=\left(\varsigma^{\prime}, \theta^{\prime}\right)$ and $z-1=z-z_{h}+u$ with $u=z_{h}-1$, one finds

$$
\begin{align*}
c^{\prime} w(z)^{\prime} & =\varsigma^{\prime} \beta_{0}^{\prime}-\left(z-z_{h}+u\right) \varsigma^{\prime} \bar{\alpha}_{0}^{\prime} A_{1} \\
& =\varsigma^{\prime}\left(\beta_{0}^{\prime}-\bar{\alpha}_{0}^{\prime} A_{1} u\right)-\left(z-z_{h}\right) \varsigma^{\prime} \alpha_{0}^{\prime} A_{1}=-\left(z-z_{h}\right) \varsigma^{\prime} \alpha_{0}^{\prime} A_{1} \tag{32}
\end{align*}
$$

where the last equality follows from (31). This shows that under the necessary and sufficient condition in Proposition 6, there is a linear combination $c^{\prime}$ of $w^{\prime}(z)$ where one can factor $z-z_{h}$ out of $c^{\prime} w^{\prime}(z)$, which reduces the order from 1 to 0 . Here $c^{\prime} w^{\prime}(z)$, which has degree equal to 1 , is replaced by $c^{\prime} w^{\prime}(z) /\left(z-z_{h}\right)=-\varsigma^{\prime} \bar{\alpha}_{0}^{\prime} A_{1}=-\lambda \varsigma^{\prime} \beta_{0}^{\prime}$, which has degree 0 . Note that from (31) the new cointegrating relation is in the span of $\beta_{0}^{\prime}$.

This can be done for all pairs $\left(\lambda, \varsigma^{\prime}\right)$. Let $\left(\lambda_{j}, \varsigma_{j}^{\prime}\right)$ be all the pairs $\left(\lambda, \varsigma^{\prime}\right)$ satisfying the assumptions of Proposition $6, j=1, \ldots, s$, and let $q^{\prime}:=\left(\lambda_{1} \varsigma_{1}, \ldots, \lambda_{k} \zeta_{s}\right)^{\prime}$. Choose also $a^{\prime}$ as some matrix $(r-s) \times r$ matrix such that $(q, a)$ is square and nonsingular; many matrices satisfy this criterion, including $q_{\perp}$. The output of Step 2 can be expressed as the following choice of $b^{(2)}(z)^{\prime}$ :

$$
b^{(2)}(z)^{\prime}=\left(\begin{array}{c}
a^{\prime} \beta_{0}^{\prime}-(z-1) a^{\prime} \bar{\alpha}_{0}^{\prime} A_{1}  \tag{33}\\
q^{\prime} \beta_{0}^{\prime} \\
\beta_{1}^{\prime}
\end{array}\right)
$$

Remark 28 ( $\mathrm{CI}(2,2)$ cointegration). This step brings out from $\phi(z)^{\prime}$ some cointegrating relations $q^{\prime} \beta_{0}^{\prime}$ that map the $I(2)$ variables directly to $I(0)$ without the help of first differences $\Delta$.
8.3. Step 3 in $I(2)$ VAR

Consider $b^{(2)}(z)^{\prime}$ in (33) and its high order coefficient matrix

$$
w_{*}^{\prime}=\left(\begin{array}{c}
-a^{\prime} \bar{\alpha}_{0}^{\prime} A_{1} \\
q^{\prime} \beta_{0}^{\prime} \\
\beta_{1}^{\prime}
\end{array}\right)
$$

Step 3 requires to find a nonzero matrix $c^{\prime}$ such that $c^{\prime} w_{*}^{\prime}=0^{\prime}$. Recall that ( $\bar{\beta}_{0}, \bar{\beta}_{1}, \bar{\beta}_{2}$ ) is square and nonsingular; hence $c^{\prime} w_{*}^{\prime}=0^{\prime}$ if and only if, partitioning $c^{\prime}$ as $c^{\prime}=\left(\zeta^{\prime}, \rho^{\prime}, \tau^{\prime}\right)$ one has

$$
0^{\prime}=c^{\prime} w_{*}^{\prime}\left(\bar{\beta}_{0}, \bar{\beta}_{1}, \bar{\beta}_{2}\right)=\left(\zeta^{\prime}, \rho^{\prime}, \tau^{\prime}\right)\left(\begin{array}{ccc}
-a^{\prime} \bar{\alpha}_{0}^{\prime} A_{1} \bar{\beta}_{0} & -a^{\prime} \bar{\alpha}_{0}^{\prime} A_{1} \bar{\beta}_{1} & -a^{\prime} \bar{\alpha}_{0}^{\prime} A_{1} \bar{\beta}_{2} \\
q^{\prime} & 0 & 0 \\
0 & I_{r_{1}} & 0
\end{array}\right)
$$

This equality can be written as

$$
\begin{align*}
\zeta^{\prime} a^{\prime} \bar{\alpha}_{0}^{\prime} A_{1}\left(\bar{\beta}_{0}, \bar{\beta}_{1}\right) & =\left(\rho^{\prime} q^{\prime}, \tau^{\prime}\right),  \tag{34}\\
\zeta^{\prime} a^{\prime} \bar{\alpha}_{0}^{\prime} A_{1} \bar{\beta}_{2} & =0 . \tag{35}
\end{align*}
$$

Remark 29 (Further degree reductions). Equation (35) requires $\zeta^{\prime}$ to be orthogonal to remaining part of the multicointegrating coefficient $a^{\prime} \bar{\alpha}_{0}^{\prime} A_{1}$ in direction of $\beta_{2}$. In addition $\zeta^{\prime}$ also needs to satisfy (34). For some configurations of dimensions, (34) could be solvable for ( $\rho^{\prime}, \tau^{\prime}$ ) in terms of other quantities; in this case (34) would not impose further restrictions.

Let also $\vartheta^{\prime}$ be any complementary matrix such that $(\zeta, \vartheta)$ is square and nonsingular; one possible choice of $\vartheta$ is $\zeta_{\perp}$. The output of Step 3 can be expressed as the following choice of $b^{(3)}(z)^{\prime}$ :

$$
b^{(3)}(z)^{\prime}=\left(\begin{array}{c}
\vartheta^{\prime} a^{\prime}\left(\beta_{0}^{\prime}-(z-1) \bar{\alpha}_{0}^{\prime} A_{1}\right)  \tag{36}\\
\zeta^{\prime} a^{\prime} \beta_{0}^{\prime} \\
q^{\prime} \beta_{0}^{\prime} \\
\beta_{1}^{\prime}
\end{array}\right)
$$

Remark 30 (Minimal basis). This step brings out from $\phi(z)^{\prime}$ some other cointegrating relations $\zeta^{\prime} a^{\prime} \beta_{0}^{\prime}$ that map the $I(2)$ variables directly to $I(0)$ without the help of first differences $\Delta$. Equation (36) shows how the canonical system of root functions can be reduced to minimal order.

Example 4 (Multicointegration coefficient in the span of $\beta_{2}$ ). Consider the special case when the multicointegrating coefficient $\bar{\alpha}_{0}^{\prime} A_{1}$ satisfies $\bar{\alpha}_{0}^{\prime} A_{1}=\bar{\alpha}_{0}^{\prime} A_{1} P_{\beta_{2}}$, i.e., it has components only in the direction of $\beta_{2}$. This special case is relevant, because $\beta_{2}^{\prime} \Delta x_{t} \sim I(1)$ while $\beta_{i}^{\prime} \Delta x_{t} \sim I(d)$ with $d \leq 0$ for $i=0,1$.

One can see that in this case the conditions in Proposition 6 are not satisfied. In fact (29) cannot hold, as $\bar{\alpha}_{0}^{\prime} A_{1} \bar{\beta}_{0}=0$. Step 2 is hence empty, and this implies that the rows including $q^{\prime}$ are missing and $a=I$ in $b^{(2)}(z)^{\prime}$ in (33) and (36).

Applying Step 3, Equation (34) is always satisfied by the choice $\rho^{\prime}=0^{\prime}, \tau^{\prime}=0^{\prime}$ because $\bar{\alpha}_{0}^{\prime} A_{1}\left(\bar{\beta}_{0}, \bar{\beta}_{1}\right)=0$. Equation (35) then reads $\zeta^{\prime} \bar{\alpha}_{0}^{\prime} A_{1} \bar{\beta}_{2}=0$, which is satisfied if and only if $\delta:=\bar{\alpha}_{0}^{\prime} A_{1} \bar{\beta}_{2}$ has reduced rank. In this case, let the rank factorization be $\delta=\psi \eta^{\prime}$, with $\psi$ and $\eta$ of full column rank. One can then let $\zeta^{\prime}=\psi_{\perp}^{\prime}$ and choose $\vartheta^{\prime}=\bar{\psi}^{\prime}$, so that

$$
b^{(3)}(z)^{\prime}=\left(\begin{array}{c}
\bar{\psi}^{\prime} \beta_{0}^{\prime}-(z-1) \eta^{\prime}  \tag{37}\\
\zeta^{\prime} \beta_{0}^{\prime} \\
\beta_{1}^{\prime}
\end{array}\right)
$$

There are several examples of this separation in the I(2) VAR literature; for example Kongsted (2005) discusses this when $r_{0}>r_{2}$.

## 9. Conclusions

This paper discusses the notion of cointegrating space for general $I(d)$ processes. The notion of cointegrating space was formally introduced in the literature by Johansen (1988) for the case of $\mathrm{I}(1)$ VAR system. The definition of the cointegrating space is simplest in the $I(1)$ case without multicointegration, because there is no need to consider vector polynomials in the lag operator.

Engle and Yoo (1991) introduced the notion of polynomial cointegrating vectors in parallel with the related one of multicointegration in Granger and Lee (1989). However, the literature has not yet discussed the notion of cointegrating space in the general polynomial case; this paper fills this gap.

In this context, this paper recognises that cointegrating vectors are in general root functions, which have been analysed at length in the mathematical and engineering literature, see e.g., Gohberg et al. (1993). This allows to characterise a number of properties of cointegrating vectors.

Canonical systems of root functions are found to provide a basis of several notions of cointegration space in the multicointegrated case. The extended local rank factorization of Franchi and Paruolo (2016) can be used to explicitly derive a canonical system of root functions. This result is constructive, as it gives an explicit way to derive such a basis from the VAR polynomial.

The canonical system of root functions constructed in this way is not necessarily of minimal polynomial degree, however. The three-step procedure of Forney (1975) to reduce this basis to minimal-degree is reviewed and restated in terms of rank factorizations. The application of this procedure to $I(2)$ VAR systems is shown to separate the polynomial and the non-polynomial cointegrating vectors.

Author Contributions: The authors contributed equally to the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.

## Appendix A

## Appendix A.1. Scalar, Vector, Matrix Analytic Functions

Consider a rational functions $a(z)$, defined as $a(z)=c(z) / d(z)$ with $c(z)$ and $d(z)$ polynomials, where $d(z) \not \equiv 0$. One can ask when $a(z)$ is analytic on $D\left(z_{\omega}, \eta\right), \eta>0$. The following remark states that this is the case provided $z_{\omega}$ is not a root of $d(z)$.

Remark A1 (Rational scalars can be analytic on $D\left(z_{\omega}, \eta\right)$ ). Let $a(z)$ be a rational function, i.e., $a(z)=c(z) / d(z)$ with $c(z)$ and $d(z)$ polynomial; assume also in addition that $d(z)$ has no root equal to $z_{\omega}$. Then $a(z)$ is analytic on $D\left(z_{\omega}, \eta\right)$, for some $\eta>0$. In fact, let $q=\operatorname{deg} d(z)$ be the degree of $d(z)$, and decompose $d(z)=\sum_{j=0}^{q} d_{j}^{\circ} z^{j}$ as $d(z)=d_{q}^{\circ} \prod_{j=1}^{n}\left(z-u_{j}\right)^{k_{j}}$, where $u_{j}$ are the roots of $d(z)$ with multiplicity $k_{j}, j=1, \ldots, n$ using the factor theorem for polynomials, see e.g., Barbeau (1989, p. 56). Then each term $\left(z-u_{j}\right)^{-k_{j}}$ has an analytic representation on $D\left(z_{\omega}, \eta\right)$, $\eta>0$, see e.g., Lemma A1 below. Note that this generates an infinite tail in a $(z)$, i.e., $a(z)$ is not polynomial in this case (unless $q=0$ ).

Lemma $\mathbf{A 1}$ (The inverse of a polynomial is analytic away form its roots). Let $u_{1}, \ldots, u_{n} \in \mathbb{C}$ be the distinct roots of a polynomial $d(z)$ with multiplicities $k_{1}, \ldots, k_{n}, k_{j} \in \mathbb{N}$, and let $v \in \mathbb{C}$ be another point, distinct from $u_{j}$; then one can pick some radius $\delta$ with $0<\delta<\min _{j=1, \ldots, n}\left|u_{j}-v\right|$ such that $d(z)^{-1}$ is analytic on $z \in D(v, \delta)$.

Proof. The polynomial $d(z)$ can be decomposed as $d(z)=a \prod_{j=1}^{n}\left(z-u_{j}\right)^{k_{j}}$. Next consider each term in the product $\left(z-u_{j}\right)^{k_{j}}$ and observe that $z-u_{j}=(z-v)-\left(u_{j}-v\right)=(v-$ $\left.u_{j}\right)\left(1-x_{j}\right)$ where $x_{j}:=(z-v) /\left(u_{j}-v\right)$. Define $0<\delta_{j}<\left|u_{j}-v\right|$ and note that $\left|x_{j}\right|<1$ for $z \in D\left(v, \delta_{j}\right)$, so that $\left(1-x_{j}\right)^{-1}=\sum_{s=0}^{\infty} x_{j}^{s}$ for $z \in D\left(v, \delta_{j}\right)$ and

$$
\left(z-u_{j}\right)^{-k_{j}}=\left(v-u_{j}\right)^{k_{j}}\left(\sum_{s=0}^{\infty}\left(\frac{z-v}{u_{j}-v}\right)^{s}\right)^{k_{j}} \quad z \in D\left(v, \delta_{j}\right) .
$$

Hence $\left(z-u_{j}\right)^{-k_{j}}$ is analytic on $z \in D\left(v, \delta_{j}\right)$ for any $j=1, \ldots, n$, and as a consequence also on $z \in D(v, \delta)$ with $0<\delta<\min _{j=1, \ldots, n}\left|u_{j}-v\right|$. This implies that $d(z)^{-1}=a^{-1} \prod_{j=1}^{n}(z-$ $\left.u_{j}\right)^{-k_{j}}$ is analytic on $z \in D(v, \delta)$.

Similarly, consider a $1 \times p$ vector function $\zeta(z)^{\prime}$ with rational entries. The denominator polynomials in all entries can be collected in a single one, the least common denominator, and hence $\zeta(z)^{\prime}$ has representation $\zeta(z)^{\prime}=\frac{1}{d(z)} b(z)^{\prime}$ where $d(z)$ is a monic polynomial and $b(z)^{\prime}$ is a $1 \times p$ vector polynomial, and $d(z)$ and $b(z)^{\prime}$ are relatively prime. The same applies to $p \times p$ rational matrix functions $C(z)$.

Example A1 (The least common denominator of bivariate rational vectors). The least common denominator can be illustrated as follows. Take a $1 \times 2$ rational row vector a $(z)^{\prime}=$ $\left(a_{1}(z), a_{2}(z)\right)=\left(c_{1}(z) / d_{1}(z), c_{2}(z) / d_{2}(z)\right)$, where $c_{i}(z), d_{i}(z)$ are (nonzero) polynomials $i=$ 1,2 ; then one can find a polynomial $d(z)$ with lowest degree such that $d(z)=h_{1}(z) d_{1}(z)=$ $h_{2}(z) d_{2}(z)$ where $h_{i}(z)$ are polynomials $i=1,2 ; d(z)$ is the least common multiple of the denominators, i.e., the least common denominator, and one has

$$
a(z)^{\prime}=\left(\frac{c_{1}(z)}{d_{1}(z)}, \frac{c_{2}(z)}{d_{2}(z)}\right)=\left(\frac{c_{1}(z) h_{1}(z)}{d(z)}, \frac{c_{2}(z) h_{2}(z)}{d(z)}\right)=: \frac{1}{d(z)} b(z)^{\prime}
$$

with $b(z)^{\prime}:=\left(b_{1}(z), b_{2}(z)\right):=\left(c_{1}(z) h_{1}(z), c_{2}(z) h_{2}(z)\right)$ where $b_{j}(z):=c_{j}(z) h_{j}(z)$ are still polynomials, so that $b(z)^{\prime}$ is a vector polynomial. The vector polynomial $b(z)^{\prime}$ and the scalar polynomial $d(z)$ are relatively prime, because there is no scalar polynomial $g(z)$ that divides both $d(z)$ and all the elements in $b(z)^{\prime}$. The polynomials in $b(z)^{\prime}$ and $d(z)$ can still be divided by a scalar in $F$, so $d(z)$ can be assumed to be monic.

Remark A2 (Rational vector and matrices). The $1 \times p$ analytic vector functions $\zeta(z)^{\prime}$ and $p \times p$ analytic matrix functions $C(z)$ can be generated as rational vectors or matrices, as long as their denominator polynomial $d(z)$ has no root equal to $z_{\omega}$. When $d(z)$ has one root equal to $z_{\omega}$ with multiplicity $m>0$, this implies that $\zeta(z)^{\prime}$ or $C(z)$ have a pole of order $m>0$ at $z_{\omega}$, and $\zeta(z)^{\prime}$ or $C(z)$ are not analytic on a disk $D\left(z_{\omega}, \eta\right)$ centered around $z_{\omega}$.

## Appendix A.2. Spans of Canonical Systems of Root Functions

This section considers linear combinations of canonical system of root functions $\phi(z)^{\prime}$ with coefficients in $F, F[z]$ and $F(z)$. Attention is first given to multiplication of a root function by a rational or polynomial scalar; next generic linear combinations of canonical system of root functions in $\phi(z)^{\prime}$ are considered.

In order to discuss results, the notion of generalized root function is introduced first.
Definition A1 (Generalized root function). Let $n \in \mathbb{Z}$ and $\zeta(z)^{\prime}$ be a root function of $C(z)$ at $z_{\omega}$ and order s, see Definition 3; then

$$
\xi(z)^{\prime}:=\left(1-z / z_{\omega}\right)^{n} \zeta(z)^{\prime}
$$

is called a generalized root function of $C(z)$ at $z_{\omega}$ with order s and exponent $n$.
Observe that this is in line with Definition 5 of generalized cointegrating vectors for rational vectors. The reason for the introduction of the notion of generalized root function is provided by the next proposition.

Proposition A1 (Multiplication by a scalar). Let $\zeta(z)^{\prime}$ be a $1 \times p$ root function for $C(z)$ of order s on $D\left(z_{\omega}, \eta\right), \eta>0$. Then
(i) if $a \in F, a \neq 0$, then $a \zeta(z)^{\prime}$ is a root function on $D\left(z_{\omega}, \eta\right)$ of order $s$;
(ii) if $a(z) \in F[z], a(z) \neq 0$, then $a(z) \zeta(z)^{\prime}$ is a generalized root function on $D\left(z_{\omega}, \eta\right)$ of order $s$ and exponent $n \in \mathbb{N}_{0}:=\mathbb{N} \cup 0$;
(iii) if $a(z) \in F(z), a(z) \neq 0$, then $a(z) \zeta(z)^{\prime}$ is a generalized root function on $D\left(z_{\omega}, \eta\right)$ of order $s$ and exponent $n \in \mathbb{Z}$.

Proof. Consider first case (iii).
(iii). $a(z)=a_{1}(z) / a_{2}(z)$ where $a_{i}(z)$ are relatively prime polynomials, $i=1,2$. If $a_{i}(z)$ has root $z_{\omega}$ then it admits representation $a_{i}(z)=\left(1-z / z_{\omega}\right)^{n_{i}} a_{i}^{\diamond}(z)$ with $n_{i} \in \mathbb{N}_{0}$ and $a_{i}^{\diamond}\left(z_{\omega}\right) \neq$ $0, i=1,2$. Hence $a(z)=a_{1}(z) / a_{2}(z)=\left(1-z / z_{\omega}\right)^{n_{1}-n_{2}} a^{\diamond}(z)$ where $a^{\diamond}(z):=a_{1}^{\diamond}(z) / a_{2}^{\diamond}(z)$ with $a^{\diamond}\left(z_{\omega}\right) \neq 0$ and $a(z) \zeta(z)^{\prime}=\left(1-z / z_{\omega}\right)^{n_{1}-n_{2}} a^{\diamond}(z) \zeta(z)^{\prime}$. The factor $\left(1-z / z_{\omega}\right)^{n_{1}-n_{2}}$ has exponent $n_{1}-n_{2}$, which can be positive, negative or 0 ; because $a_{i}(z)$ are relatively prime polynomials, $i=1,2$, either $n_{1}>0$ or $n_{2}>0$ or $n_{1}=n_{2}=0$. The factor $a^{\diamond}(z) \zeta(z)^{\prime}$ is a generalized root function of order $s$, because $\zeta(z)^{\prime}$ is a root function of order $s$ and the scalar factor $a^{\diamond}(z)$ satisfies $a^{\diamond}\left(z_{\omega}\right) \neq 0$, so that $a^{\diamond}\left(z_{\omega}\right) \zeta\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$. This shows that $a(z) \zeta(z)^{\prime}$ is a generalized root function of order $s$ and exponent $n=n_{1}-n_{2}$.
(ii). Set $a_{2}(z)=1$ in the proof of $\left.i i i\right)$, and note that the exponent is $n_{1}$, which is either 0 or positive.
(i). Set $a_{1}(z)=a, a_{2}(z)=1$ in the proof of $\left.i i i\right)$, and note that the exponent is $n_{1}=0$.

Remark A3 (A generalized root function is meromorphic). A generalized root function $\xi(z)^{\prime}$ is analytic on $D\left(z_{\omega}, \eta\right)$ except for the possibility to have poles at the isolated point $z_{\omega}$, i.e., it is a meromorphic function on $D\left(z_{\omega}, \eta\right)$.

Remark A4 (A generalized root function can be analytic). When the exponent $n$ of $\xi(z)^{\prime}$ is zero, the generalized root function $\xi(z)^{\prime}$ coincides with the root function $\zeta(z)^{\prime}$. When the exponent $n$ of $\xi(z)^{\prime}$ is positive, then the generalized root function $\xi(z)^{\prime}$ has a zero at $z_{\omega}$. In both cases $\xi(z)^{\prime}$ is analytic. So a generalized root function can be analytic (with or without a zero at $z_{\omega}$ ).

Remark A5 (Generalized root function and cointegration). Observe that Definition A1 implies the following: given a meromorphic function $\xi(z)^{\prime}$, check if it has a root or a pole at $z_{\omega}$; this function is a generalized root functions if, after removing the pole or the zero at $z_{\omega}$ by multiplying it by $\left(1-z / z_{\omega}\right)^{-n}$ where $n$ is the order of the root or of the pole, the resulting function is a root function, i.e., a cointegrating vector. This is in line with Definition 5.

Attention is now turned to linear combinations of a canonical system of root functions $\phi(z)^{\prime}$. The scalars of the linear combination can be in $F, F[z]$ or $F(z)$. The main result in Proposition A2 below is that $F[z]$-linear combinations of $\phi(z)^{\prime}$ generate a generalized root function possibly with a zero at $z_{\omega}$, while $F(z)$-linear combinations of $\phi(z)^{\prime}$ generate a generalized root function possibly with a pole or a zero at $z_{\omega}$.

In the following, let $v^{\prime}=\left(v_{1}, \ldots, v_{m}\right)^{\prime} \in F^{m}$ be a $1 \times m$ vector with elements in $F$. Let also $A_{v}$ be the set of non-zero entries in $v, A_{v}:=\left\{i: v_{i} \neq 0\right\}$, with $n_{v}$ the cardinality of $A_{v}$ and $\left(i_{1}, \ldots, i_{n_{v}}\right)$ the ordered set of indices in $A_{v}, i_{1}<\cdots<i_{n_{v}}, i_{j} \in A_{v}$. Similarly, let $w(z)^{\prime}=\left(w_{1}(z), \ldots, w_{m}(z)\right)^{\prime} \in F[z]^{m}$ be a $1 \times m$ vector with polynomial elements in $w_{i}(z) \in F[z]$ with $\left(j_{1}, \ldots, j_{n_{w}}\right)$ its ordered set of indices of nonzero elements in $A_{w}:=\left\{i: w_{i}(z) \neq 0\right\}$, and let finally $u(z)^{\prime}=\left(u_{1}(z), \ldots, u_{m}(z)\right)^{\prime} \in F(z)^{m}$ be a $1 \times m$ vector with rational elements in $u_{i}(z) \in F(z)$ with $\left(k_{1}, \ldots, k_{n_{u}}\right)$ as its ordered set of indices of nonzero elements in $A_{u}:=\left\{i: u_{i}(z) \neq 0\right\}$.

Proposition A2 (Linear combinations). Let $\phi(z)^{\prime}=\left(\phi_{1}(z), \ldots \phi_{m}(z)\right)^{\prime}$ be a canonical system of root functions of $C(z)$ on a disc $D\left(z_{\omega}, \eta\right), \eta>0$ with orders $s_{1}, \ldots, s_{m}$; let also $v^{\prime} \in F^{m}$, $w(z)^{\prime} \in F[z]^{m}$ and $u(z)^{\prime} \in F(x)^{m}$ be nonzero vectors; one has:
(i) $v^{\prime} \phi(z)^{\prime}=\sum_{i=1}^{m} v_{i} \phi_{i}(z)^{\prime}$ is a root function of order $s=\min _{i \in A_{v}} s_{i}$;
(ii) $w(z)^{\prime} \phi(z)^{\prime}=\sum_{j=1}^{m} w_{j}(z) \phi_{j}(z)^{\prime}$ is a generalized root function, with exponent $q=\min _{j \in A_{w}}$ $\left(q_{j}\right) \geq 0$ where $q_{j}$ is the order of $z_{\omega}$ as a zero of $w_{j}(z)$, and with order $s:=\min _{j \in A_{i v}}\left(q_{j}-q+\right.$ $\left.s_{j}\right)>0 ;$
(iii) $u(z)^{\prime} \phi(z)^{\prime}=\sum_{k=1}^{m} u_{k}(z) \phi_{k}(z)^{\prime}$ is a generalized root function, possibly with a pole or a zero at $z_{\omega}$, with exponent $q=\min _{k \in A_{u}}\left(q_{k}\right) \in \mathbb{Z}$ where $q_{k}$ is the order of $z_{\omega}$ as a pole or as a zero of $u_{k}(z)$, and with order $s=\min _{k \in A_{u}}\left(q_{k}-q+s_{k}\right)>0$.

Proof. (i). By definition $\phi_{i}(z)^{\prime}=\sum_{j=0}^{s_{i}}\left(z-z_{\omega}\right)^{j} \phi_{i j}^{\prime}$, analytic on $D\left(z_{\omega}, \eta\right)$ and $\phi_{i j}^{\prime} \in F^{p}$. One finds $\sum_{i=1}^{m} v_{i} \phi_{i}(z)^{\prime}=\sum_{j=0}^{s_{i}}\left(z-z_{\omega}\right)^{j} \varphi_{j}^{\prime}$ with $\varphi_{j}^{\prime}:=\sum_{i=1}^{m} v_{i} \phi_{i j}^{\prime} \in F^{p}$ because $F$ is a field, and hence it is closed under multiplication. Hence $v^{\prime} \phi(z)^{\prime}$ is a polynomial with coefficients vectors in $F^{p}$, of the same form as each $\phi_{i}(z)^{\prime}$, and one finds that

$$
\begin{equation*}
v^{\prime} \phi(z)^{\prime} C(z)=\sum_{i=1}^{m} v_{i} \phi_{i}(z)^{\prime} C(z)=\sum_{i \in A_{v}} v_{i}\left(z-z_{\omega}\right)^{s_{i}} \widetilde{\phi}_{i}(z)^{\prime}=\left(z-z_{\omega}\right)^{s} \widetilde{\phi}(z)^{\prime} \tag{A1}
\end{equation*}
$$

where $s:=\min \left\{s_{i_{1}}, \ldots, s_{i_{v}}\right\}, \widetilde{\phi}(z)^{\prime}:=\sum_{h=1}^{n_{v}}\left(z-z_{\omega}\right)^{s_{h}-s} v_{i_{h}} \widetilde{\phi}_{i_{h}}(z)^{\prime}$ and $\widetilde{\phi}_{i_{h}}\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$. Note that because $v^{\prime}$ is a nonzero vector, the set $A_{v}$ is not empty. Next observe that $\widetilde{\phi}\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$ otherwise this would contradict the property of $\phi_{i_{h}}(z)^{\prime}$ to be of maximal order and linearly independent from the previous root function $\phi_{i}(z)^{\prime}$ for $i<i_{h}$. This shows that $v^{\prime} \phi(z)^{\prime}$ is a root function of order $s$.
(ii). Consider $w(z)^{\prime} \phi(z)^{\prime}=\sum_{i=1}^{m} w_{i}(z) \phi_{i}(z)^{\prime}$, where by Proposition A1. (i), one has that $w_{i}(z) \phi_{i}(z)^{\prime}$ is a generalized root function with representation $w_{i}(z) \phi_{i}(z)^{\prime}=(1-$ $\left.z / z_{\omega}\right)^{q_{i}} w_{i}^{\diamond}(z) \phi_{i}(z)^{\prime}$ say, with $q_{i} \geq 0$ and $w_{i}^{\diamond}(z) \phi_{i}(z)^{\prime}$ a root function of order $s_{i}$. Let $q:=$ $\min \left(q_{j_{1}}, \ldots, q_{j_{w}}\right)$, and note that $w(z)^{\prime} \phi(z)^{\prime}=\left(1-z / z_{\omega}\right)^{9} \zeta(z)^{\prime}$ with $\zeta(z)^{\prime}:=\sum_{h=1}^{n_{w w}}(1-$ $\left.z / z_{\omega}\right)^{q_{j_{h}}-q} w_{j_{h}}^{\diamond}(z) \phi_{j_{h}}(z)^{\prime}$. In order to show that $\zeta\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$, let $B_{w}$ be the set of indices $j \in A_{w}$ with $q_{j}=q$, and observe that $\zeta\left(z_{\omega}\right)^{\prime}=\sum_{j \in B_{w}} w_{j}^{\diamond}\left(z_{\omega}\right) \phi_{j}\left(z_{\omega}\right)^{\prime}$ where $w_{j}^{\diamond}\left(z_{\omega}\right) \neq 0$ by construction and $\phi_{j}\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$ by the definition of root function. If $\zeta\left(z_{\omega}\right)^{\prime}=0^{\prime}$ this would imply that there is a nonzero linear combination of $\phi\left(z_{\omega}\right)^{\prime}$ equal to $0^{\prime}$, i.e., that $\phi\left(z_{\omega}\right)^{\prime}$ is not of full row rank, which contradicts the construction in Definition 4. This implies that $\zeta\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$, and that $w(z)^{\prime} \phi(z)^{\prime}$ is a generalized root function of order $q$.

Next, because $\phi_{j}(z)^{\prime}$ is a root function of order $s_{j}$ one has

$$
\begin{aligned}
\zeta(z)^{\prime} C(z) & =\sum_{h=1}^{n_{w}}\left(1-z / z_{\omega}\right)^{q_{j_{h}}-q} w_{j_{h}}^{\diamond}(z) \phi_{j_{h}}(z)^{\prime} C(z) \\
& =\sum_{h=1}^{n_{w v}}\left(1-z / z_{\omega}\right)^{q_{j_{h}}-q+s_{j_{h}}} w_{j_{h}}^{\diamond}(z) \widetilde{\phi}_{j_{h}}(z)^{\prime}=\left(1-z / z_{\omega}\right)^{s} \widetilde{\zeta}(z)^{\prime}
\end{aligned}
$$

where $\widetilde{\zeta}(z)^{\prime}:=\sum_{h=1}^{n_{w}}\left(1-z / z_{\omega}\right)^{q_{j_{h}}-q+s_{j_{h}}-s} w_{j_{h}}^{\diamond}(z) \widetilde{\phi}_{j_{h}}(z)^{\prime}$. Finally, in order to prove that the order of the generalized root function is $s$, one needs to show that $\widetilde{\zeta}\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$. Let $C_{w}$ be the set of indices $j \in A_{w}$ with $q_{j_{h}}-q+s_{j_{h}}=s$, and observe that $\widetilde{\zeta}\left(z_{\omega}\right)^{\prime}:=\sum_{j \in C_{w}} w_{j}^{\diamond}\left(z_{\omega}\right) \widetilde{\phi}_{j}\left(z_{\omega}\right)^{\prime}$ where $w_{j}^{\diamond}\left(z_{\omega}\right) \neq 0$ and $\phi_{j}\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$ as above. If $\widetilde{\zeta}\left(z_{\omega}\right)^{\prime}=0^{\prime}$, then there exists a nonzero linear combination of $\phi\left(z_{\omega}\right)^{\prime}$ equal to $0^{\prime}$, which would imply the existence of a root function of higher order obtained by combination of the rows in $\phi(z)^{\prime}$ with index $C_{w}$, which contradict the fact that the orders are chosen to be maximal. This implies that the order of the generalized root function is equal to $s$.
(iii). The proof is the same as in $i i$ ). Note that here $q_{i}$ may be negative.

Remark A6 (Closure with respect to linear combinations). Proposition A2 shows that $F[z]$ linear combinations and $F(z)$-linear combinations of a canonical systems of root functions $\phi(z)^{\prime}$ produce generalized root functions. Note that $\phi(z)^{\prime}$ is itself a set of generalized root functions (with 0 exponent). Hence, in this sense, generalized root functions are closed under $F[z]$-linear combinations and $F(z)$-linear combinations.

Remark A7 (Spans). Indicate the set of G-linear combinations $\phi(z)^{\prime}$ as $\operatorname{row}_{G}\left(\phi(z)^{\prime}\right)$, where $G=F, F(z), F[z]$. It is simple to observe that

$$
\begin{equation*}
\operatorname{row}_{F}\left(\phi(z)^{\prime}\right) \subset \operatorname{row}_{F[z]}\left(\phi(z)^{\prime}\right) \subset \operatorname{row}_{F(z)}\left(\phi(z)^{\prime}\right) \tag{A2}
\end{equation*}
$$

Remark A8 (Role of characteristics of canonical system of root functions). The proof of Proposition A2 reveals that, in order to conclude that a $F[z]$ - or $F(z)$-linear combination of $\phi(z)^{\prime}$ is a generalized root function, the property that $\phi\left(z_{\omega}\right)^{\prime}$ is of full row rank plays a crucial role. In fact, when reaching the equality $w(z)^{\prime} \phi(z)^{\prime}=\left(1-z / z_{\omega}\right)^{9} \zeta(z)^{\prime}$ where $q$ is the exponent of the linear combination, one can show that $\zeta\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$ by making use of this property only, without using the maximal orders of the root functions in $\phi(z)^{\prime}$. This proves the following corollary.

Corollary A1 (Linear combinations of a set of root functions). Replace the canonical system of root functions $\phi(z)^{\prime}$ in Proposition $A 2$ with a set of $m$ root functions $\xi(z)^{\prime}$ for $C(z)$ on $D\left(z_{\omega}, \eta\right)$, $\eta>0$ such that $\xi\left(z_{\omega}\right)^{\prime}$ is of full row rank; then the $F[z]$ - or $F(z)$-linear combinations $w(z)^{\prime} \xi(z)^{\prime}$ and $u(z)^{\prime} \xi(z)^{\prime}$ are generalized root functions with the same exponents as in Proposition A2.

## Appendix A.3. Truncations of Root Functions

This section discusses how the truncation of a root function still delivers a root function, possibly of lower order. The main implication of this property is that one can take any element in $\operatorname{row}_{G}\left(\phi(z)^{\prime}\right)$ for $G=F, F[z], F(z)$ and obtain other root functions by truncation, thus enlarging the set of root functions that can be generated from $\operatorname{row}_{G}\left(\phi(z)^{\prime}\right)$.

Let $\zeta(z)^{\prime}:=\sum_{j=0}^{\infty}\left(z-z_{\omega}\right)^{j} \zeta_{j}^{\prime}$ be a $1 \times p$ root function of order $s$ of $C(z)$ on $D\left(z_{\omega}, \eta\right)$, $\eta>0$, and indicate the truncation of $\zeta(z)^{\prime}$ to a polynomial of degree $r$ as $\zeta^{(r)}(z)^{\prime}:=$ $\sum_{j=0}^{r}\left(z-z_{\omega}\right)^{j} \zeta_{j}^{\prime}$; the remainder $\zeta(z)^{\prime}-\zeta^{(r)}(z)^{\prime}=\sum_{j=r+1}^{\infty}\left(z-z_{\omega}\right)^{j} \zeta_{j}^{\prime}$ is called the tail of $\zeta(z)^{\prime}$. The following proposition clarifies that one can modify the tail of a root function without affecting its property to factor some power of $\left(1-z / z_{\omega}\right)$ from $C(z)$. One special case is that one can delete the tail after the order $s$ of the root function without changing its order.

Proposition A3 (Truncations). Let $\zeta(z)^{\prime}:=\sum_{j=0}^{\infty}\left(z-z_{\omega}\right)^{j} \zeta_{j}^{\prime}$ be a root function of order sfor $C(z)=\sum_{j=0}^{\infty}\left(z-z_{\omega}\right)^{j} C_{j}$ on $D\left(z_{\omega}, \eta\right), \eta>0, \zeta\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$, and let $\psi(z)^{\prime}:=\sum_{j=0}^{\infty}\left(z-z_{\omega}\right)^{j} \psi_{j}^{\prime}$ be a $1 \times p$ vector function, analytic on $D\left(z_{\omega}, \eta\right)$; then
(i) for an integer $\ell \geq 1$, the $1 \times p$ row vector $\xi(z)^{\prime}$ with

$$
\begin{equation*}
\xi(z)^{\prime}:=\zeta(z)^{\prime}+\left(z-z_{\omega}\right)^{\ell} \psi(z)^{\prime} \tag{A3}
\end{equation*}
$$

is still a root function on $D\left(z_{\omega}, \eta\right)$ of order $n \geq \min (\ell, s)$;
(ii) if one chooses $\ell \leq s$, in the definition (A3) of $\xi(z)^{\prime}$ with $\psi(z)^{\prime}$ proportional to the tail of $\zeta(z)^{\prime}$, a special case of $i$ ) is that the truncation of $\zeta(z)^{\prime}$ to the polynomial $\zeta^{(\ell)}(z)^{\prime}:=\sum_{j=0}^{\ell}\left(z-z_{\omega}\right)^{j} \zeta_{j}^{\prime}$ of order $\ell$ is also a root function on $D\left(z_{\omega}, \eta\right)$ of order $n \geq \ell$;
(iii) finally if $\zeta(z)^{\prime}:=\sum_{j=0}^{\infty}\left(z-z_{\omega}\right)^{j} \zeta_{j}^{\prime}$ is a root function of order s in a canonical system of root functions of $C(z)$ at $z_{\omega}$, then its truncation $\zeta^{(s-1)}(z)^{\prime}:=\sum_{j=0}^{s-1}\left(z-z_{\omega}\right)^{j} \zeta_{j}^{\prime}$ to a polynomial of degree $s-1$ is still a root function of $C(z)$ at $z_{\omega}$ on $D\left(z_{\omega}, \eta\right)$ of order $s$.

Proof. (i). By definition one has $\zeta(z)^{\prime} C(z)=\left(z-z_{\omega}\right)^{s} \widetilde{\zeta}(z)^{\prime}$ with $\widetilde{\zeta}\left(z_{\omega}\right)^{\prime}=\sum_{h=0}^{s} \zeta_{h}^{\prime} C_{s-h} \neq$ $0^{\prime}$. Hence, setting $q:=\min (\ell, s)$, one finds

$$
\widetilde{\zeta}(z)^{\prime} C(z)=\zeta(z)^{\prime} C(z)+\left(z-z_{\omega}\right)^{\ell} \psi(z)^{\prime} C(z)=\left(z-z_{\omega}\right)^{q} \widetilde{\xi}(z)^{\prime} .
$$

where $\widetilde{\xi}(z)^{\prime}=\left(z-z_{\omega}\right)^{s-q} \widetilde{\zeta}(z)^{\prime}+\left(z-z_{\omega}\right)^{\ell-q} \psi(z)^{\prime} C(z)$. If $\widetilde{\xi}\left(z_{\omega}\right)^{\prime} \neq 0^{\prime}$, then $\tilde{\xi}(z)^{\prime}$ is a root function of order $q$. If, instead, $\widetilde{\xi}\left(z_{\omega}\right)^{\prime}=0^{\prime}$, then $\xi^{\prime}(z)^{\prime}$ is a root function of order $n$ greater than $q$; in any case $n \geq q$, with $n$ finite by Proposition 2 .
(ii). Choose $\psi(z)^{\prime}=\left(z-z_{\omega}\right)^{-\ell}\left(\zeta(z)^{\prime}-\zeta^{(\ell)}(z)^{\prime}\right)=\sum_{j=\ell+1}^{\infty}\left(z-z_{\omega}\right)^{j-\ell} \zeta_{j}^{\prime}$ in (A3), so that $\xi(z)^{\prime}=\zeta^{(\ell)}(z)^{\prime}$. The statement follows from $(i)$.
(iii). The coefficients $\zeta(z)^{\prime}$ and $\zeta^{(s-1)}(z)^{\prime}$ generate the same coefficients $W_{h}:=\sum_{j=0}^{h} \zeta_{j}^{\prime} C_{h-j}$ for $h=0, \ldots, s-1$ in the convolution $\zeta(z)^{\prime} C(z)=W(z):=\sum_{h=0}^{\infty}\left(z-z_{\omega}\right)^{h} W_{h}$, where $W_{h}=0$ for $h=0, \ldots, s-1$ by definition of order $s$, see (8). This implies that $\zeta^{(s-1)}(z)^{\prime}$ is a root function at least of order $s$. However, because root functions in a canonical system of root functions are chosen of maximal order, the order of $\zeta^{(s-1)}(z)^{\prime}$ is equal to $s$. This completes the proof.

Remark A9 (Truncated cointegrating vectors). Proposition A3. (ii) implies that truncating a cointegrating vector to order $\ell<s$ preserves the cointegrating property, but not necessarily the order s.

Remark A10 (Cointegrating vectors in $I_{\omega}(1)$ VAR can be chosen not to be polynomial). Consider Example 1, where the orders of integration of (polynomial) linear combinations can be either 1 or 0 . In this case, root function are of order at most $s=1$, and Proposition A3. (iii) ensures that the root functions can be truncated to order s-1=0, i.e., to non-polynomial linear combinations.

Remark A11 (A generic $I_{\omega}(1)$ process may have polynomial cointegration relations). Consider now the generic case of an $I(1)$ process. The orders of integration of (polynomial) linear combinations can be $0,-1,-2, \cdots-d$ say, with $d>0$. In this case, root function are of order at most $s=1,2 \ldots, d+1$, and Proposition A3. (iii) ensures that the root functions can be truncated to order $d$. If $d>0$ this may require polynomial linear combinations also in the $I_{\omega}(1)$ case.

Remark A12 (Polynomial cointegration vectors in $I_{\omega}(2)$ VAR can be chosen of order at most 1). Consider Example 2, where the orders of integration of (polynomial) linear combinations can be either 2,1 or 0 . In this case, root function are of order at most $s=2$, and Proposition A3. (iii) ensures that the root functions can be truncated to order $s-1=1$, i.e., to polynomial linear combinations of order 1 .

Remark A13 (Multicointegrated systems require polynomial cointegration relations). As shown in the previous three remarks, in general, multicointegrated systems require to consider polynomial linear combinations.

## Notes

$1 \quad$ See Engle and Granger (1987, pp. 253-54). Here $N$ in their notation is replaced by $p$ and $\alpha$ with $\beta$ for consistency with the rest of the paper.
2 The following notation is employed: $F=\mathbb{R}, \mathbb{C}$ indicates either the field of real numbers $\mathbb{R}$ or the field of complex numbers $\mathbb{C}$ and if a matrix $A=\left(a_{1}, \ldots, a_{n}\right)$ is written in terms of its columns, $\operatorname{col}_{F}(A)$ indicates the column span of $A$ with coefficients in $F$, i.e., $\operatorname{col}_{F}(A):=\left\{v: v=\sum_{j=1}^{n} c_{i} a_{i}, c_{i} \in F\right\}$ and $\operatorname{row}_{F}\left(A^{\prime}\right)$ denotes the row span of $A^{\prime}$ with coefficients in $F$, i.e., $\operatorname{row}_{F}\left(A^{\prime}\right):=\left\{v^{\prime}:\right.$ $\left.v^{\prime}=\sum_{j=1}^{n} c_{i} a_{i}^{\prime}, c_{i} \in F\right\}$, where $A^{\prime}$ indicates the conjugate transpose of $A$. Hence $v \in \operatorname{col}_{F}(A)$ if and only if $v^{\prime} \in \operatorname{row}_{F}\left(A^{\prime}\right)$, i.e., the spaces coincide but the former contains column vectors while the latter contains row vectors. Here the row form is employed.
$3 \quad \varepsilon_{t}$ could be taken to be non-autocorrelated instead of i.i.d. with no major changes in the results in the paper.
4 This result is usually stated as $x_{t}=u_{t}-a_{0}$ where $a_{0}:=x_{0}-u_{0}$ is a generic constant, see e.g., Hannan and Deistler (1988) Equation (1.2.15).
5 In fact, substituting $C(z)=C+\widetilde{C}(z)(1-z)$, one finds $\zeta^{\prime} \Delta x_{t}=\zeta^{\prime} C \varepsilon_{t}+\zeta^{\prime} \widetilde{C}(L) \Delta \varepsilon_{t}$, and applying $\mathcal{S}$ to both sides gives $\zeta^{\prime} x_{t}-$ $\zeta^{\prime} x_{0}=\zeta^{\prime} \mathcal{C S} \varepsilon_{t}+u_{t}-u_{0}$ where $u_{t}:=\zeta^{\prime} \widetilde{C}(L) \varepsilon_{t}$ is stationary. The term $\mathcal{S} \varepsilon_{t}$ is a bilateral random walk (Franchi and Paruolo 2019), a nonstationary process, so that the l.h.s. can be made stationary if and only if the coefficient $\zeta^{\prime} \mathrm{C}$ loading $\mathcal{S} \varepsilon_{t}$ is 0 .
6 Theorem 3 provides two constructions of the local Smith factorization.
7 In this case $A(z)$ is analytic for all $z \in \mathbb{C}$.
8 In the first sentence in Definition 3.1 of Franchi and Paruolo (2019) ' $r_{0}^{\max }:=p-r_{0}{ }^{\prime}$ should read ' $r_{0}^{\max }:=p^{\prime}$. The results of Franchi and Paruolo (2019, Theorem 3.3) are applied setting $F(z)$ there equal to $A(z)$ here.

9 The present statement follows by Franchi and Paruolo (2019, Theorem 3.5) with $F(z)$ and $\Phi(z)$ there equal to $A(z)$ and $\Xi(z)^{-1}$ here.

## References

Barbeau, Edward J. 1989. Polynomials. Berlin and Heidelberg: Springer.
Bauer, Dietmar, and Martin Wagner. 2012. A State Space Canonical Form for Unit Root Processes. Econometric Theory 28: 1313-49.
Beare, Brendan, and Won-ki Seo. 2020. Representation of $\mathrm{I}(1)$ and $\mathrm{I}(2)$ autoregressive Hilbertian processes. Econometric Theory 36: 773-802.
Beare, Brendan, Juwon Seo, and Won-ki Seo. 2017. Cointegrated Linear Processes in Hilbert Space. Journal of Time Series Analysis 38: 1010-27.
Engle, Robert F., and Clive W. J. Granger. 1987. Co-integration and Error Correction: Representation, Estimation, and Testing. Econometrica 55: 251-76.
Engle, Robert F., and Sam B. Yoo. 1991. Cointegrated economic time series: An overview with new results. In Long-Run Economic Relations: Readings in Cointegration. Edited by Robert Engle and Clive Granger. Oxford: Oxford University Press, pp. 237-66.
Engsted, Tom, and Søren Johansen. 2000. Granger's Representation Theorem and Multicointegration. In Cointegration, Causality and Forecasting: Festschrift in Honour of Clive Granger. Edited by R. F. Engle and H. White. Oxford: Oxford University Press, pp. 200-12.
Forney, G. David, Jr. 1975. Minimal bases of rational vector spaces, with applications to multivariable linear systems. SIAM Journal on Control 13: 493-520. [CrossRef]
Franchi, Massimo, and Paolo Paruolo. 2011. Inversion of regular analytic matrix functions: Local Smith form and subspace duality. Linear Algebra and Its Applications 435: 2896-912.
Franchi, Massimo, and Paolo Paruolo. 2016. Inverting a matrix function around a singularity via local rank factorization. SIAM Journal of Matrix Analysis and Applications 37: 774-97.
Franchi, Massimo, and Paolo Paruolo. 2019. A general inversion theorem for cointegration. Econometric Reviews 38: 1176-201. [CrossRef]
Franchi, Massimo, and Paolo Paruolo. 2020. Cointegration in functional autoregressive processes. Econometric Theory 36: 803-39.
Gohberg, Israel, Marinus A. Kaashoek, and Frederik Van Schagen. 1993. On the local theory of regular analytic matrix functions. Linear Algebra and Its Applications 182: 9-25.
Granger, Clive W. J., and Tae-Hwy Lee. 1989. Investigation of production, sales and inventory relationships using multicointegration and non-symmetric error correction models. Journal of Applied Econometrics 4: S145-59.
Gregoir, Stèphane M. 1999. Multivariate time series with various hidden unit roots, Part I. Econometric Theory 15: 435-68.
Hannan, Edward J., and Manfred Deistler. 1988. The Statistical Theory of Linear Systems. Hoboken: John Wiley \& Sons.
Howlett, Phil G. 1982. Input retrieval in finite dimensional linear systems. Journal of Australian Mathematical Society (Series B) 23: 357-82.
Hungerford, Thomas W. 1980. Algebra. Berlin and Heidelberg: Springer.
Hylleberg, Svend, Robert F. Engle, Clive W. J. Granger, and Sam B. Yoo. 1990. Seasonal integration and cointegration. Journal of Econometrics 44: 215-38.
Johansen, Søren. 1988. Statistical Analysis of Cointegration Vectors. Journal of Economic Dynamics and Control 12: 231-54.
Johansen, Søren. 1991. Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. Econometrica 59: 1551-80.
Johansen, Søren. 1992. A representation of vector autoregressive processes integrated of order 2. Econometric Theory 8: 188-202.
Johansen, Søren. 1995. Identifying restrictions of linear equations with applications to simultaneous equations and cointegration. Journal of Econometrics 69: 111-32.
Johansen, Søren. 1996. Likelihood-Based Inference in Cointegrated Vector Auto-Regressive Models. Oxford: Oxford University Press.
Johansen, Søren, and E. Schaumburg. 1998. Likelihood analysis of seasonal cointegration. Journal of Econometrics 88: 301-39.
Kongsted, Hans Christian. 2005. Testing the nominal-to-real transformation. Journal of Econometrics 124: 205-25. [CrossRef]
Mosconi, Rocco, and Paolo Paruolo. 2017. Identification conditions in simultaneous systems of cointegrating equations with integrated variables of higher order. Journal of Econometrics 198: 271-76. [CrossRef]
Seo, Won-ki. 2019. Cointegration and Representation of Cointegrated Autoregressive Processes in Banach Spaces. arXiv arXiv:1712.08748v4.

## Article

# Johansen's Reduced Rank Estimator Is GMM 

Bruce E. Hansen<br>Department of Economics, University of Wisconsin, Madison, WI 53706, USA; bruce.hansen@wisc.edu; Tel.: +1-608-263-3880

Received: 30 January 2018; Accepted: 16 May 2018; Published: 18 May 2018


#### Abstract

The generalized method of moments (GMM) estimator of the reduced-rank regression model is derived under the assumption of conditional homoscedasticity. It is shown that this GMM estimator is algebraically identical to the maximum likelihood estimator under normality developed by Johansen (1988). This includes the vector error correction model (VECM) of Engle and Granger. It is also shown that GMM tests for reduced rank (cointegration) are algebraically similar to the Gaussian likelihood ratio tests. This shows that normality is not necessary to motivate these estimators and tests.


Keywords: GMM; VECM; reduced rank
JEL Classification: C3

## 1. Introduction

The vector error correction model (VECM) of Engle and Granger (1987) is one of the most widely used time-series models in empirical practice. The predominant estimation method for the VECM is the reduced-rank regression method introduced by Johansen (1988, 1991, 1995). Johansen's estimation method is widely used because it is straightforward, it is a natural extension of the VAR model of Sims (1980), and it is computationally tractable.

Johansen motivated his estimator as the maximum likelihood estimator (MLE) of the VECM under the assumption that the errors are i.i.d. normal. For many users, it is unclear whether the estimator has a broader justification. In contrast, it is well known that least-squares estimation is both maximum likelihood under normality and method of moments under uncorrelatedness.

This paper provides the missing link. It is shown that Johansen's reduced-rank estimator is algebraically identical to the generalized method of moments (GMM) estimator of the VECM, under the imposition of conditional homoscedasticity. This GMM estimator only uses uncorrelatedness and homoscedasticity. Thus Johansen's reduced-rank estimator can be motivated under much broader conditions than normality.

The asymptotic efficiency of the estimator in the GMM class relies on the assumption of homoscedasticity (but not normality). When homoscedasticity fails, the reduced-rank estimator loses asymptotic efficiency but retains its interpretation as a GMM estimator.

It is also shown that the GMM tests for reduced (cointegration) rank are nearly identical to Johansen's likelihood ratio tests. Thus the standard likelihood ratio tests for cointegration can be interpreted more broadly as GMM tests.

This paper does not introduce new estimation nor inference methods. It merely points out that the currently used methods have a broader interpretation than may have been understood. The results leave open the possibility that new GMM methods that do not impose homoscedasticity could be developed.

This connection is not new. In a different context, Adrian et al. (2015) derived the equivalence of the likelihood and minimum-distance estimators of the reduced-rank model. The equivalence
between the Limited Information Maximum Likelihood (LIML) estimator (which has a dual relation with reduced-rank regression) and a minimum distance estimator was discovered by Goldberger and Olkin (1971). Recently, Kolesár (2018) drew out connections between likelihood-based and minimum-distance estimation of endogenous linear regression models.

This paper is organized as follows. Section 2 introduces reduced-rank regression models and Johansen's estimator. Section 3 presents the GMM and states the main theorems demonstrating the equivalence of the GMM and MLE. Section 4 presents the derivation of the GMM estimator. Section 5 contains two technical results relating generalized eigenvalue problems and the extrema of quadratic forms.

## 2. Reduced-Rank Regression Models

The VECM for $p$ variables of cointegrating rank $r$ with $k$ lags is

$$
\begin{equation*}
\Delta X_{t}=\alpha \beta^{\prime} X_{t-1}+\sum_{i=1}^{k-1} \Gamma_{i} \Delta X_{t-i}+\Phi D_{t}+e_{t} \tag{1}
\end{equation*}
$$

where $D_{t}$ are the deterministic components. Observations are $t=1, \ldots, T$. The matrices $\alpha$ and $\beta$ are $p \times r$ with $r \leq p$. This is a famous workhorse model in applied time series, largely because of the seminal work of Engle and Granger (1987).

The primary estimation method for the VECM is known as reduced-rank regression and was developed by Johansen $(1988,1991,1995)$. Algebraically, the VECM (1) is a special case of the reduced-rank regression model:

$$
\begin{equation*}
Y_{t}=\alpha \beta^{\prime} X_{t}+\Psi Z_{t}+e_{t} \tag{2}
\end{equation*}
$$

where $Y_{t}$ is $p \times 1, X_{t}$ is $m \times 1$, and $Z_{t}$ is $q \times 1$. The coefficient matrix $\alpha$ is $p \times r$ and $\beta$ is $m \times r$ with $r \leq \min (m, p)$. Johansen derived the MLE for model (2) under the assumption that $e_{t}$ is i.i.d. $N(0, \Omega)$. This immediately applies to the VECM (1) and is the primary application of reduced-rank regression in econometrics.

Canonical correlations were introduced by Hotelling (1936), and reduced-rank regression was introduced by Bartlett (1938). A complete theory was developed by Anderson and Rubin $(1949,1950)$ and Anderson (1951). These authors developed the MLE for the model:

$$
\begin{align*}
Y_{t} & =\Pi X_{t}+e_{t}  \tag{3}\\
\Gamma^{\prime} \Pi & =0 \tag{4}
\end{align*}
$$

where $\Gamma$ is $p \times(p-r)$ and is unknown. This is an alternative parameterization of (2) without the covariates $Z_{t}$. Anderson and Rubin $(1949,1950)$ considered the case $p-r=1$ and primarily focused on estimation of the vector $\Gamma$. Anderson (1951) considered the case $p-r \geq 1$.

While the models (2) and (3)-(4) are equivalent and thus have the same MLE, the different parameterizations led the authors to different derivations. Anderson and Rubin derived the estimator of (3) and (4) by a tedious application of constrained optimization. (Specifically, they maximized the likelihood of (3) imposing the constraint (4) using Lagrange multiplier methods. The solution turned out to be tedious because (4) is a nonlinear function of the parameters $\Gamma$ and $\Pi$.) The derivation is so cumbersome that it is excluded from nearly all statistics and econometrics textbooks, despite the fact that it is the source of the famous LIML estimator.

The elegant derivation used by Johansen (1988) is algebraically unrelated to that of Anderson-Rubin and is based on applying a concentration argument to the product structure in (2). It is similar to the derivation in Tso (1981), although the latter did not include the covariates $Z_{t}$. Johansen's derivation is algebraically straightforward and thus is widely taught to students.

It is useful to briefly describe the likelihood problem. The log-likelihood for model (2) under the assumption that $e_{t}$ is i.i.d. $N(0, \Omega)$ is

$$
\begin{equation*}
\ell(\alpha, \beta, \Psi, \Omega)=-\frac{T}{2} \log \operatorname{det} \Omega-\frac{1}{2} \sum_{t=1}^{T}\left(Y_{t}-\alpha \beta^{\prime} X_{t}-\Psi Z_{t}\right)^{\prime} \Omega^{-1}\left(Y_{t}-\alpha \beta^{\prime} X_{t}-\Psi Z_{t}\right) \tag{5}
\end{equation*}
$$

The MLE maximizes $\ell(\alpha, \beta, \Psi, \Omega)$. Johansen's solution is as follows. Define the projection matrix $M_{Z}=I_{T}-Z\left(Z^{\prime} Z\right)^{-1} Z^{\prime}$ and the residual matrices $\widetilde{Y}=M_{Z} Y$ and $\widetilde{X}=M_{Z} X$. Consider the generalized eigenvalue problem:

$$
\begin{equation*}
\left|\widetilde{X}^{\prime} \widetilde{Y}\left(\widetilde{Y}^{\prime} \widetilde{Y}\right)^{-1} \widetilde{Y}^{\prime} \widetilde{X}-\widetilde{X}^{\prime} \widetilde{X} \lambda\right|=0 \tag{6}
\end{equation*}
$$

The solutions $1>\hat{\lambda}_{1}>\cdots>\hat{\lambda}_{p}>0$ satisfy

$$
\tilde{X}^{\prime} \tilde{Y}\left(\tilde{Y}^{\prime} \widetilde{Y}\right)^{-1} \tilde{Y}^{\prime} \tilde{X} v_{i}=\tilde{X}^{\prime} \tilde{X} \widehat{v}_{i} \widehat{\lambda}_{i}
$$

where $\left(\widehat{\lambda}_{i}, \widehat{v}_{i}\right)$ are known as the generalized eigenvalues and eigenvectors of $\widetilde{X}^{\prime} \widetilde{Y}\left(\widetilde{Y}^{\prime} \widetilde{Y}\right)^{-1} \widetilde{Y}^{\prime} \widetilde{X}$ with respect to $\widetilde{X}^{\prime} \widetilde{X}$. The normalization $\widehat{v}_{i}^{\prime} \widetilde{X}^{\prime} \widetilde{X} \widehat{v}_{i}=1$ is imposed.

Given the normalization $\beta^{\prime} \widetilde{X}^{\prime} \widetilde{X} \beta=I_{r}$, Johansen's reduced-rank estimator for $\beta$ is

$$
\widehat{\beta}_{\mathrm{mle}}=\left[\widehat{v}_{1}, \ldots, \widehat{v}_{r}\right] .
$$

The MLE $\widehat{\alpha}_{\text {mle }}$ and $\widehat{\Psi}_{\text {mle }}$ are found by least-squares regression of $Y_{t}$ on $\widehat{\beta}_{\text {mle }}^{\prime} X_{t}$ and $Z_{t}$.

## 3. Generalized Method of Moments

Define $W_{t}=\left(X_{t}^{\prime}, Z_{t}^{\prime}\right)^{\prime}$. The GMM estimator of the reduced-rank regression model (2) is derived under the standard orthogonality restriction:

$$
\begin{equation*}
\mathbb{E}\left(W_{t} e_{t}^{\prime}\right)=0 \tag{7}
\end{equation*}
$$

plus the homoscedasticity condition:

$$
\begin{equation*}
\mathbb{E}\left(e_{t} e_{t}^{\prime} \otimes W_{t} W_{t}^{\prime}\right)=\Omega \otimes Q \tag{8}
\end{equation*}
$$

where $\Omega=\mathbb{E}\left(e_{t} e_{t}^{\prime}\right)$ and $Q=\mathbb{E}\left(W_{t} W_{t}^{\prime}\right)$. These moment conditions are implied by the normal regression model. (Equations (7) and (8) can be deduced from the first-order conditions for maximization of (5)). Because (7) and (8) can be deduced from (5) but not vice versa, the moment condition model (7) and (8) is considerably more general than the normal regression model (5).

The efficient GMM criterion (see Hansen 1982) takes the form

$$
J_{r}(\alpha, \beta, \Psi)=T \bar{g}_{r}(\alpha, \beta, \Psi)^{\prime} \widehat{V}^{-1} \bar{g}_{r}(\alpha, \beta, \Psi)
$$

where

$$
\begin{align*}
\bar{g}_{r}(\alpha, \beta, \Psi) & =\frac{1}{T} \sum_{t=1}^{n}\left(\left(Y_{t}-\alpha \beta^{\prime} X_{t}-\Psi Z_{t}\right) \otimes W_{t}\right)  \tag{9}\\
\widehat{V} & =\widehat{\Omega} \otimes \widehat{Q} \\
\widehat{\Omega} & =\frac{1}{T} \sum_{t=1}^{n} \widehat{e}_{t} \hat{e}_{t}^{\prime}  \tag{10}\\
\widehat{Q} & =\frac{1}{T} \sum_{t=1}^{n} W_{t} W_{t}^{\prime}
\end{align*}
$$

and $\widehat{e}_{t}$ are the least-squares residuals of the unconstrained model:

$$
\widehat{e}_{t}=Y_{t}-\widehat{\Pi} X_{t}-\widehat{\Psi} Z_{t}
$$

The GMM estimator are the parameters that jointly minimize the criterion $J_{r}(\alpha, \beta, \Psi)$ subject to the normalization $\beta^{\prime} \widetilde{X}^{\prime} \widetilde{X} \beta=I_{r}$ :

$$
\left(\widehat{\alpha}_{\mathrm{gmm}}, \widehat{\beta}_{\mathrm{gmm}}, \widehat{\Psi}_{\mathrm{gmm}}\right)=\underset{\beta^{\prime} \tilde{X}^{\prime} \widetilde{\mathrm{X}} \beta=I_{r}}{\operatorname{argmin}} J_{r}(\alpha, \beta, \Psi)
$$

The main contribution of the paper is the following surprising result.
Theorem 1. $\left(\widehat{\alpha}_{\mathrm{gmm}}, \widehat{\beta}_{\mathrm{gmm}}, \widehat{\Psi}_{\mathrm{gmm}}\right)=\left(\widehat{\alpha}_{\text {mle }}, \widehat{\beta}_{\text {mle }}, \widehat{\Psi}_{\text {mle }}\right)$.
Theorem 2. $J_{r}\left(\widehat{\alpha}_{\mathrm{gmm}}, \widehat{\beta}_{\mathrm{gmm}}, \widehat{\Psi}_{\mathrm{gmm}}\right)=\operatorname{tr}\left(\widehat{\Omega}^{-1}\left(\widetilde{Y}^{\prime} \widetilde{Y}\right)\right)-T p-T \sum_{i=1}^{r} \frac{\widehat{\lambda}_{i}}{1-\widehat{\lambda}_{i}}$, where $\widehat{\lambda}_{i}$ are the eigenvalues from (6).

Theorem 1 states that the GMM estimator is algebraically identical to the Gaussian maximum likelihood estimator.

This shows that Johansen's reduced-rank regression estimator is not tied to the normality assumption. This is similar to the equivalence of least-squares as a method of moments estimator and the Gaussian MLE in the regression context.

The key is the use of the homoscedastic weight matrix. This shows that the Johansen reduced-rank estimator is an efficient GMM estimator under conditional homoscedasticity. When homoscedasticity fails, the Johansen reduced-rank estimator continues to be a GMM estimator but is no longer the efficient GMM estimator.

It is important to understand that Theorem 1 is different from the trivial statement that the MLE is GMM applied to the first-order condition of the likelihood (e.g., Hall (2005), Section 3.8.1). Specifically, if you take the derivatives of the Gaussian log-likelihood function (5) and treat these as moment conditions and solve, this is a GMM estimator, and thus MLE can be interpreted as GMM. That is not what Theorem 1 states.

GMM hypothesis tests can be constructed by the difference in the GMM criteria; tests for reduced rank are considered, which in the context of VECM are tests for cointegration rank. The model

$$
Y_{t}=\Pi X_{t}+\Psi Z_{t}+e_{t}
$$

is taken and the following hypotheses on reduced rank are considered:

$$
\mathbb{H}_{r}: \operatorname{rank}(\Pi)=r
$$

The GMM test statistic for $\mathbb{H}_{r}$ against $\mathbb{H}_{r+1}$ is

$$
C_{r, r+1}=\min _{\beta^{\prime} \widetilde{X}^{\prime} \widetilde{\mathrm{X}} \beta=I_{r}} J_{r}(\alpha, \beta, \Psi)-\min _{\beta^{\prime} \widetilde{X}^{\prime} \widetilde{\mathrm{X}} \beta=I_{r+1}} J_{r+1}(\alpha, \beta, \Psi) .
$$

The GMM test statistic for $\mathbb{H}_{r}$ against $\mathbb{H}_{p}$ is

$$
C_{r, p}=\min _{\beta^{\prime} \tilde{X}^{\prime} \tilde{X} \beta=I_{r}} J_{r}(\alpha, \beta, \Psi)-\min _{\beta^{\prime} \tilde{X}^{\prime} \tilde{X} \beta=I_{p}} J_{p}(\alpha, \beta, \Psi) .
$$

Theorem 3. The GMM test statistics for reduced rank are

$$
\begin{aligned}
C_{r, r+1} & =T\left(\frac{\hat{\lambda}_{r+1}}{1-\widehat{\lambda}_{r+1}}\right), \\
C_{r, p} & =T \sum_{i=r+1}^{p} \frac{\hat{\lambda}_{i}}{1-\hat{\lambda}_{i}}
\end{aligned}
$$

where $\widehat{\lambda}_{i}$ are the eigenvalues from (6).
Here it is recalled in contrast that the likelihood ratio test statistics derived by Johansen are

$$
\begin{aligned}
L R_{r, r+1} & =-T \log \left(1-\widehat{\lambda}_{r+1}\right) \\
L R_{r, p} & =-T \sum_{i=r+1}^{p} \log \left(1-\widehat{\lambda}_{r+1}\right)
\end{aligned}
$$

The GMM test statistic $C_{r, r+1}$ and the likelihood ratio (LR) statistic $L R_{r, r+1}$ yield equivalent tests, as they are monotonic functions of one another. (If the bootstrap is used to assess significance, the two statistics will yield numerically identical $p$-values.) They are asymptotically identical under standard approximations and in practice will be nearly identical, because the eigenvalues $\widehat{\lambda}_{i}$ tend to be quite small in value (at least under the null hypothesis), so that $-\log (1-\lambda) \approx \lambda /(1-\lambda) \approx \lambda$. For $p-(r+1)>1$, the GMM test statistic $C_{r, p}$ and the LR statistic $L R_{r, p}$ do not provide equivalent tests (they cannot be written as monotonic functions of one another), but they are also asymptotically equivalent and will be nearly identical in practice.

An interesting connection noted by a referee is that the statistic $C_{r, p}$ was proposed by Pillai (1955) and Muirhead (1982, Section 11.2.8).

## 4. Derivation of the GMM Estimator

It is convenient to rewrite the criterion in standard matrix notation, defining the matrices $Y, X, Z$, and $W$ by stacking the observations. Model (2) is

$$
Y=X \beta \alpha^{\prime}+Z \Psi^{\prime}+e
$$

The moment (9) is

$$
\bar{g}_{r}(\alpha, \beta, \Psi)=\frac{1}{T} \operatorname{vec}\left(W^{\prime}\left(Y-X \beta \alpha^{\prime}-Z \Psi^{\prime}\right)\right)
$$

Using the relation

$$
\operatorname{tr}(A B C D)=\operatorname{vec}\left(D^{\prime}\right)^{\prime}\left(C^{\prime} \otimes A\right) \operatorname{vec}(B)
$$

the following is obtained:

$$
\begin{aligned}
J_{r}(\alpha, \beta, G) & =\operatorname{Tg}_{r}(\alpha, \beta, \Psi)^{\prime}\left(\widehat{\Omega}^{-1} \otimes \widehat{Q}^{-1}\right) \bar{g}_{r}(\alpha, \beta, \Psi) \\
& =\operatorname{vec}\left(W^{\prime}\left(Y-X \beta \alpha^{\prime}-Z \Psi^{\prime}\right)\right)^{\prime}\left(\widehat{\Omega}^{-1} \otimes\left(W^{\prime} W\right)^{-1}\right) \operatorname{vec}\left(W^{\prime}\left(Y-X \beta \alpha^{\prime}-Z \Psi^{\prime}\right)\right) \\
& =\operatorname{tr}\left(\widehat{\Omega}^{-1}\left(Y-X \beta \alpha^{\prime}-Z \Psi^{\prime}\right)^{\prime} W\left(W^{\prime} W\right)^{-1} W^{\prime}\left(Y-X \beta \alpha^{\prime}-Z \Psi^{\prime}\right)\right)
\end{aligned}
$$

Following the concentration strategy used by Johansen, $\beta$ is fixed and $\alpha$ and $\Psi$ are concentrated out, producing a concentrated criterion that is a function of $\beta$ only. The system is linear in the regressors $X \beta$ and Z. Given the homoscedastic weight matrix, the GMM estimator of $(\alpha, \Psi)$ is multivariate least-squares. Using the partialling out (residual regression) approach, the least-squares residual can
be written as the residual from the regression of $\widetilde{Y}$ on $\widetilde{X} \beta$, where $\widetilde{Y}=M_{Z} Y$ and $\widetilde{X}=M_{Z} X$ are the residuals from regressions on $Z$. That is, the least-squares residual is

$$
\begin{aligned}
\widehat{e}(\beta) & =\widetilde{Y}-\widetilde{X} \beta\left(\beta^{\prime} \widetilde{X}^{\prime} \widetilde{X} \beta\right)^{-1} \beta^{\prime} \widetilde{X}^{\prime} \widetilde{Y} \\
& =\widetilde{Y}-\widetilde{X} \beta \beta^{\prime} \widetilde{X}^{\prime} \widetilde{Y}
\end{aligned}
$$

where the second equality uses the normalization $\beta^{\prime} \widetilde{X}^{\prime} \widetilde{X} \beta=I_{r}$. Because the space spanned by $W=(X, Z)$ equals that spanned by $(\widetilde{X}, Z)$, the following can be written:

$$
W\left(W^{\prime} W\right)^{-1} W^{\prime}=Z\left(Z^{\prime} Z\right)^{-1} Z^{\prime}+\widetilde{X}\left(\widetilde{X}^{\prime} \widetilde{X}\right)^{-1} \widetilde{X}^{\prime}
$$

Because $Z^{\prime} \widehat{e}(\beta)=0$, then

$$
\begin{aligned}
W\left(W^{\prime} W\right)^{-1} W^{\prime} \widehat{e}(\beta) & =\widetilde{X}\left(\widetilde{X}^{\prime} \widetilde{X}\right)^{-1} \widetilde{X}^{\prime} \widehat{e}(\beta) \\
& =\widetilde{X}\left(\widetilde{X}^{\prime} \widetilde{X}\right)^{-1} \widetilde{X}^{\prime} \widetilde{Y}-\widetilde{X} \beta \beta^{\prime} \widetilde{X}^{\prime} \widetilde{Y}
\end{aligned}
$$

and

$$
\begin{aligned}
\widehat{e}(\beta)^{\prime} W\left(W^{\prime} W\right)^{-1} W^{\prime} \widehat{e}(\beta) & =\widetilde{Y}^{\prime} \widetilde{X}\left(\widetilde{X}^{\prime} \widetilde{X}\right)^{-1} \widetilde{X}^{\prime} \widetilde{Y}-\widetilde{Y}^{\prime} \widetilde{X} \beta \beta^{\prime} \widetilde{X}^{\prime} \widetilde{Y} \\
& =\widetilde{Y}^{\prime} \widetilde{Y}-\widetilde{Y}^{\prime} M_{\widetilde{X}} \widetilde{Y}-\widetilde{Y}^{\prime} \widetilde{X} \beta \beta^{\prime} \widetilde{X}^{\prime} \widetilde{Y}
\end{aligned}
$$

where

$$
M_{\tilde{X}}=I-\widetilde{X}\left(\widetilde{X}^{\prime} \widetilde{X}\right)^{-1} \widetilde{X}^{\prime}
$$

Using the partialling out (residual regression) approach, the variance estimator (10) can be written as

$$
\widehat{\Omega}=\frac{1}{T} Y^{\prime}\left(I-W\left(W^{\prime} W\right)^{-1} W^{\prime}\right) Y=\frac{1}{T} \widetilde{Y}^{\prime} M_{\tilde{X}} \widetilde{Y}
$$

Thus the concentrated GMM criterion is

$$
\begin{align*}
J_{r}^{*}(\beta) & =\operatorname{tr}\left(\widehat{\Omega}^{-1} \widehat{e}(\beta)^{\prime} W\left(W^{\prime} W\right)^{-1} W^{\prime} \widehat{e}(\beta)\right) \\
& =\operatorname{tr}\left(\widehat{\Omega}^{-1}\left(\widetilde{Y}^{\prime} \widetilde{Y}\right)\right)-\operatorname{tr}\left(\widehat{\Omega}^{-1}\left(\widetilde{Y}^{\prime} M_{\tilde{X}^{\prime}} \widetilde{Y}\right)\right)-\operatorname{tr}\left(\widehat{\Omega}^{-1}\left(\widetilde{Y}^{\prime} \widetilde{X} \beta \beta^{\prime} \widetilde{X}^{\prime} \widetilde{Y}\right)\right) \\
& =\operatorname{tr}\left(\widehat{\Omega}^{-1}\left(\widetilde{Y}^{\prime} \widetilde{Y}\right)\right)-T p-T \operatorname{tr}\left(\beta^{\prime} \widetilde{X}^{\prime} \widetilde{Y}\left(\widetilde{Y}^{\prime} M_{\tilde{X}} \widetilde{Y}\right)^{-1} \widetilde{Y}^{\prime} \widetilde{X} \beta\right) \tag{11}
\end{align*}
$$

The GMM estimator minimizes $J_{r}^{*}(\beta)$ or, equivalently, maximizes the third term in (11). This is a generalized eigenvalue problem. Lemma 2 (in the next section) shows that the solution is $\widehat{\beta}_{\mathrm{gmm}}=\left[\widetilde{v}_{1}, \ldots, \widetilde{v}_{r}\right]$ as claimed.

Because the estimates $\widehat{\alpha}_{\text {gmm }}$ and $\widehat{\Psi}_{\text {gmm }}$ are found by regression given $\widehat{\beta}_{\text {gmm }}$, and because this is equivalent with the MLE, it is also concluded that $\widehat{\alpha}_{\mathrm{gmm}}=\widehat{\alpha}_{\text {mle }}$ and $\widehat{\Psi}_{\mathrm{gmm}}=\widehat{\Psi}_{\text {mle }}$. This completes the proof of Theorem 1.

To establish Theorem 2, Lemma 2 also shows that the minimum of the criterion is

$$
\begin{aligned}
J_{r}\left(\widehat{\alpha}_{\mathrm{gmm}}, \widehat{\beta}_{\mathrm{gmm}}, \widehat{\Psi}_{\mathrm{gmm}}\right) & =\min _{\beta^{\prime} \widetilde{X}^{\prime} \widetilde{X} \beta=I_{r}} J_{r}(\alpha, \beta, G) \\
& =\min _{\beta^{\prime} \widetilde{X}^{\prime} \widetilde{X}^{\prime} \beta=I_{r}} J_{r}^{*}(\beta) \\
& =\operatorname{tr}\left(\widehat{\Omega}^{-1}\left(\widetilde{Y}^{\prime} \widetilde{Y}\right)\right)-T p-T \max _{\beta^{\prime} \widetilde{X}^{\prime} \widetilde{X} \beta=I_{r}} \operatorname{tr}\left(\beta^{\prime} \widetilde{X}^{\prime} \widetilde{Y}\left(\widetilde{Y}^{\prime} M_{\widetilde{X}^{\prime}} \widetilde{Y}\right)^{-1} \widetilde{Y}^{\prime} \widetilde{X} \beta\right) \\
& =\operatorname{tr}\left(\widehat{\Omega}^{-1}\left(\widetilde{Y}^{\prime} \widetilde{Y}\right)\right)-T p-T \sum_{i=1}^{r} \frac{\widehat{\lambda}_{i}}{1-\widehat{\lambda}_{i}}
\end{aligned}
$$

This establishes Theorem 2.

## 5. Extrema of Quadratic Forms

To establish Theorems 1 and 2, a simple extrema property is necessary. First, a simple property that relates the maximization of quadratic forms to generalized eigenvalues and eigenvectors is given. It is a slight extension of Theorem 11.13 of Magnus and Neudecker (1988).

Lemma 1. Suppose $A$ and $C$ are $p \times p$ real symmetric matrices with $C>0$. Let $\lambda_{1}>\cdots>\lambda_{p}>0$ be the generalized eigenvalues of $A$ with respect to $C$ and $v_{1}, \ldots, v_{p}$ be the associated eigenvectors. Then

$$
\max _{\beta^{\prime} \subset \beta=I_{r}} \operatorname{tr}\left(\beta^{\prime} A \beta\right)=\sum_{i=1}^{r} \lambda_{i}
$$

and

$$
\underset{\beta^{\prime} \subset \beta=I_{r}}{\operatorname{argmax}} \operatorname{tr}\left(\beta^{\prime} A \beta\right)=\left[v_{1}, \ldots, v_{r}\right] .
$$

Proof. Define $\gamma=C^{1 / 2 \prime} \beta$ and $\bar{A}=C^{-1 / 2} A C^{-1 / 2 \prime}$. The eigenvalues of $\bar{A}$ are equal to the generalized eigenvalues $\lambda_{i}$ of $A$ with respect to $C$. The associated eigenvectors of $\bar{A}$ are $C^{1 / 2 \prime} v_{i}$. Thus by Theorem 11.13 of Magnus and Neudecker (1988),

$$
\max _{\beta^{\prime} \subset \beta=I_{r}} \operatorname{tr}\left(\beta^{\prime} A \beta\right)=\max _{\gamma^{\prime} \gamma=I_{r}} \operatorname{tr}\left(\gamma^{\prime} \bar{A} \gamma\right)=\sum_{i=1}^{r} \lambda_{i}
$$

and

$$
\begin{aligned}
\underset{\beta^{\prime} C \beta=I_{r}}{\operatorname{argmax}} \operatorname{tr}\left(\beta^{\prime} A \beta\right) & =C^{-1 / 2 \prime} \underset{\gamma^{\prime} \gamma=I_{r}}{\operatorname{argmax}} \operatorname{tr}\left(\gamma^{\prime} \bar{A} \gamma\right) \\
& =C^{-1 / 2 \prime} C^{1 / 2 \prime}\left[v_{1}, \ldots, v_{r}\right] \\
& =\left[v_{1}, \ldots, v_{r}\right]
\end{aligned}
$$

as claimed.
Lemma 2. Let $M_{X}=I-X\left(X^{\prime} X\right)^{-1} X^{\prime}$. If $X^{\prime} X>0$ and $Y^{\prime} M_{X} Y>0$ then

$$
\max _{\beta^{\prime} X^{\prime} X \beta=I_{r}} \operatorname{tr}\left(\beta^{\prime} X^{\prime} Y\left(Y^{\prime} M_{X} Y\right)^{-1} Y^{\prime} X \beta\right)=\sum_{i=1}^{r} \frac{\lambda_{i}}{1-\lambda_{i}}
$$

and

$$
\underset{\beta^{\prime} X^{\prime} X \beta=I_{r}}{\operatorname{argmax}} \operatorname{tr}\left(\beta^{\prime} X^{\prime} Y\left(Y^{\prime} M_{X} Y\right)^{-1} Y^{\prime} X \beta\right)=\left[v_{1}, \ldots, v_{r}\right],
$$

where $1>\lambda_{1}>\cdots>\lambda_{p}>0$ are the generalized eigenvalues of $X^{\prime} Y\left(Y^{\prime} Y\right)^{-1} Y^{\prime} X$ with respect to $X^{\prime} X$, and $v_{1}, \ldots, v_{p}$ are the associated eigenvectors.

Proof. By Lemma 1,

$$
\max _{\beta^{\prime} X^{\prime} X \beta=I_{r}} \operatorname{tr}\left(\beta^{\prime} X^{\prime} Y\left(Y^{\prime} M_{X} Y\right)^{-1} Y^{\prime} X \beta\right)=\sum_{i=1}^{r} \widetilde{\lambda}_{i}
$$

and

$$
\underset{\beta^{\prime} X^{\prime} X \beta=I_{r}}{\operatorname{argmax}} \operatorname{tr}\left(\beta^{\prime} X^{\prime} Y\left(Y^{\prime} M_{X} Y\right)^{-1} Y^{\prime} X \beta\right)=\left[\widetilde{v}_{1}, \ldots, \widetilde{v}_{r}\right],
$$

where $\widetilde{\lambda}_{1}>\cdots>\tilde{\lambda}_{p}>0$ are the generalized eigenvalues of $X^{\prime} Y\left(Y^{\prime} M_{X} Y\right)^{-1} Y^{\prime} X$ with respect to $X^{\prime} X$ and $\widetilde{\nu}_{1}, \ldots, \widetilde{v}_{p}$ are the associated eigenvectors. The proof is established by showing that $\widetilde{\lambda}_{i}=\lambda_{i} /\left(1-\lambda_{i}\right)$ and $\widetilde{v}_{i}=v_{i}$.

Let $(\widetilde{v}, \widetilde{\lambda})$ be a generalized eigenvector/eigenvalue pair of $X^{\prime} Y\left(Y^{\prime} M_{X} Y\right)^{-1} Y^{\prime} X$ with respect to $X^{\prime} X$. The pair satisfies

$$
\begin{equation*}
X^{\prime} Y\left(Y^{\prime} M_{X} Y\right)^{-1} Y^{\prime} X \widetilde{v}=X^{\prime} X \tilde{v} \widetilde{\lambda} \tag{12}
\end{equation*}
$$

By the Woodbury matrix identity (e.g., Magnus and Neudecker (1988), Equation (7)),

$$
\begin{aligned}
\left(Y^{\prime} M_{X} Y\right)^{-1} & =\left(Y^{\prime} Y-Y^{\prime} X\left(X^{\prime} X\right)^{-1} X^{\prime} Y\right)^{-1} \\
& =\left(Y^{\prime} Y\right)^{-1}+\left(Y^{\prime} Y\right)^{-1} Y^{\prime} X\left(X^{\prime} X-X^{\prime} Y\left(Y^{\prime} Y\right)^{-1} Y^{\prime} X\right)^{-1} X^{\prime} Y\left(Y^{\prime} Y\right)^{-1} \\
& =\left(Y^{\prime} Y\right)^{-1}+\left(Y^{\prime} Y\right)^{-1} Y^{\prime} X\left(X^{\prime} M_{Y} X\right)^{-1} X^{\prime} Y\left(Y^{\prime} Y\right)^{-1}
\end{aligned}
$$

where $M_{Y}=I-Y\left(Y^{\prime} Y\right)^{-1} Y^{\prime}$. Thus

$$
\begin{aligned}
X^{\prime} Y\left(Y^{\prime} M_{X} Y\right)^{-1} Y^{\prime} X & =X^{\prime} Y\left(Y^{\prime} Y\right)^{-1} Y^{\prime} X+X^{\prime} Y\left(Y^{\prime} Y\right)^{-1} Y^{\prime} X\left(X^{\prime} M_{Y} X\right)^{-1} X^{\prime} Y\left(Y^{\prime} Y\right)^{-1} Y^{\prime} X \\
& =X^{\prime} P_{Y} X+X^{\prime} P_{Y} X\left(X^{\prime} M_{Y} X\right)^{-1} X^{\prime} P_{Y} X \\
& =X^{\prime} X\left(X^{\prime} M_{Y} X\right)^{-1} X^{\prime} P_{Y} X
\end{aligned}
$$

where $P_{Y}=Y\left(Y^{\prime} Y\right)^{-1} Y^{\prime}$ and the final equality uses $X^{\prime} P_{Y} X=X^{\prime} X-X^{\prime} M_{Y} X$. Substituting into (12) produces

$$
X^{\prime} X\left(X^{\prime} M_{Y} X\right)^{-1} X^{\prime} P_{Y} X \widetilde{v}=X^{\prime} X \widetilde{v} \widetilde{\lambda}
$$

Multiplying both sides by $\left(X^{\prime} M_{Y} X\right)\left(X^{\prime} X\right)^{-1}$, this implies

$$
\begin{aligned}
X^{\prime} P_{Y} X \widetilde{v} & =X^{\prime} M_{Y} X \widetilde{v} \widetilde{\lambda} \\
& =X^{\prime} X \widetilde{v} \widetilde{\lambda}-X^{\prime} P_{Y} X \widetilde{v} \widetilde{\lambda}
\end{aligned}
$$

By collecting terms,

$$
X^{\prime} P_{Y} X \widetilde{v}(1+\widetilde{\lambda})=X^{\prime} X \widetilde{v} \widetilde{\lambda}
$$

which implies

$$
X^{\prime} P_{Y} X \widetilde{v}=X^{\prime} X \widetilde{v} \frac{\widetilde{\lambda}}{(1+\widetilde{\lambda})}
$$

This is an eigenvalue equation. It shows that $\widetilde{\lambda} /(1+\widetilde{\lambda})=\lambda$ is a generalized eigenvalue and $\widetilde{v}$ is the associated eigenvector of $X^{\prime} P_{Y} X$ with respect to $X^{\prime} X$. Solving, $\widetilde{\lambda}=\lambda /(1-\lambda)$. This means that the generalized eigenvalues of $X^{\prime} Y\left(Y^{\prime} M_{X} Y\right)^{-1} Y^{\prime} X$ with respect to $X^{\prime} X$ are $\lambda_{i} /\left(1-\lambda_{i}\right)$ and $v_{i}$. Because $\lambda /(1-\lambda)$ is monotonically increasing on $[0,1)$ and $\lambda_{i}<1$, it follows that the orderings of $\lambda_{i}$ and $\widetilde{\lambda}_{i}$ are identical. Thus $\widetilde{\lambda}_{i}=\lambda_{i} /\left(1-\lambda_{i}\right)$ as claimed.

Acknowledgments: This research is supported by the National Science Foundation and the Phipps Chair. Thanks to Richard Crump, the co-editors, and two referees for helpful comments on an earlier version. The author gives special thanks to Soren Johansen and Katerina Juselius for many years of stunning research, stimulating conversations, and impeccable scholarship.
Conflicts of Interest: The author declares no conflict of interest.

## References

Adrian, Tobias, Richard K. Crump, and Emanuel Moench. 2015. Regression-based estimation of dynamic asset pricing models. Journal of Financial Economics 118: 211-44. [CrossRef]
Anderson, Theodore Wilbur. 1951. Estimating linear restrictions on regression coefficeints for multivariate normal distributions. Annals of Mathematical Statistics 22: 327-50. [CrossRef]
Anderson, Theodore Wilbur, and Herman Rubin. 1949. Estimation of the parameters of a single equation in a complete system of stochastic equations. Annals of Mathematical Statistics 20: 46-63. [CrossRef]
Anderson, Theodore Wilbur, and Herman Rubin. 1950. The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. Annals of Mathematical Statistics 21: 570-82. [CrossRef]
Bartlett, Maurice S. 1938. Further aspects of the theory of multiple regression. Proceedings of the Cambridge Philosophical Society 34: 33-40. [CrossRef]
Engle, Robert F., and Clive W. J. Granger. 1987. Co-integration and error correction: Representation, estimation, and testing. Econometrica 55: 251-76. [CrossRef]
Goldberger, Arthur S., and Ingram Olkin. 1971. A minimum-distance interpretation of limited-information estimation. Econometrica 39: 635-49. [CrossRef]
Hall, Alastair R. 2005. Generalized Method of Moments. Oxford: Oxford University Press.
Hansen, Lars Peter. 1982. Large sample properties of generalized method of moments estimators. Econometrica 50: 1029-54. [CrossRef]
Hotelling, Harold. 1936. Relations between two sets of variates. Biometrika 28: 321-77. [CrossRef]
Johansen, Søren. 1988. Statistical analysis of cointegration vectors. Journal of Economic Dynamics and Control 12: 231-54. [CrossRef]
Johansen, Søren. 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica 59: 1551-80. [CrossRef]
Johansen, Søren. 1995. Likelihood-Based Inference in Cointegrated Vector Auto-Regressive Models. Oxford: Oxford University Press.
Kolesár, Michal. 2018. Minimum distance approach to inference with many instruments. Journal of Econometrics 204: 86-100. [CrossRef]
Magnus, Jan R., and Heinz Neudecker. 1988. Matrix Differential Calculus with Applications in Statistics and Econometrics. New York: Wiley.
Muirhead, Robb J. 1982. Aspects of Multivariate Statistical Theory. New York: Wiley.
Pillai, K. C. S. 1955. Some new test criteria in multivariate analysis. The Annals of Mathematical Statistics 26: 117-21. [CrossRef]
Sims, Christopher A. 1980. Macroeconomics and reality. Econometrica 48: 1-8. [CrossRef]
Tso, M. K.-S. 1981. Reduced-rank regression and canonical analysis. Journal of the Royal Statistical Society, Series B 43: 183-89.

# Article <br> Asymptotic Theory for Cointegration Analysis When the Cointegration Rank Is Deficient 

David H. Bernstein ${ }^{1}$ and Bent Nielsen ${ }^{2, *}$<br>1 Department of Economics, University of Miami, Coral Gables, FL 33146, USA; dbernstein@bus.miami.edu<br>2 Department of Economics \& Nuffield College \& Programme on Economic Modelling, University of Oxford, Oxford OX1 1NF, UK<br>* Correspondence: bent.nielsen@nuffield.ox.ac.uk

Received: 3 May 2018; Accepted: 8 January 2019; Published: 18 January 2019


#### Abstract

We consider cointegration tests in the situation where the cointegration rank is deficient. This situation is of interest in finite sample analysis and in relation to recent work on identification robust cointegration inference. We derive asymptotic theory for tests for cointegration rank and for hypotheses on the cointegrating vectors. The limiting distributions are tabulated. An application to US treasury yields series is given.


Keywords: cointegration; rank deficiency; weak identification
JEL Classification: C32

## 1. Introduction

Determination of the cointegration rank is an important part of analyzing the cointegrated vector autoregressive model in the framework of Johansen (1988, 1991, 1995), Johansen and Juselius (1990), and Juselius (2006). We consider the rank deficient case where the cointegration rank of the data generating process is smaller than the rank used in the statistical analysis. In that case, the data generating process has more unit roots than the number of unit roots imposed in the statistical analysis and the usual asymptotic theory fails. We provide asymptotic theory for cointegration rank tests and tests on cointegration vectors along with simulated tables of the asymptotic distributions.

Cointegration analysis is conducted in three steps. First, the specification of the model is checked. Second, the rank is determined using a sequential procedure using Dickey-Fuller type distributions. Third, the cointegrating vectors are estimated and restrictions can be tested using standard inference. Asymptotic theory shows that estimated rank is consistent in the sense that the probability that the estimated rank is not equal to the true rank equals the size of tests, whereas the probability that the estimated rank is too small vanishes, see Johansen $(1992,1995)$ and Paruolo $(2001)$. Hence, the rank deficiency problem does not arise in the asymptotic analysis.

In practice, rank deficiency matters in two ways. The asymptotic theory often suffers from considerable finite sample distortion. Further, if an investigator wants to focus on the inference on the cointegrating relations then problems can arise if the rank is taken as known when in fact it is deficient. These problems mirror those of instrumental variable estimation with weak instruments, see Mavroeidis et al. (2014).

When conducting inference on the cointegrating vector under near rank deficiency the parameters are weakly identified. At the extreme when testing on the cointegrating vector in the case of a deficient rank the model is mis-specified. This problem arises in cointegration as well as in instrumental variable estimation. In both cases maximum likelihood is conducted using reduced rank regression. The weak identification problem has attracted considerable attention in the instrumental variable literature, see for instance Mavroeidis et al. (2014). Khalaf and Urga (2014) discussed the weak
identification problem for cointegration, that is when testing for a known cointegrating vector in the nearly rank deficient situation. These authors investigate various methods to adjust the asymptotic distribution in the weak identification case. This includes a bounds-based critical value suggested by Dufour (1997). This method requires knowledge of the asymptotic theory for the rank deficient case, which we provide here.

The practical problem of ignoring rank deficiency is illustrated using yield curve data. The expectation hypothesis is often interpreted as follows. Interest rates at different maturities are integrated series, but cointegrate so that spreads are stationary. Spreads are often found to be non-stationary. Thus, it is quite possible that a pair of interest rates do not cointegrate. An investigator may proceed by assuming cointegration when there is none, so that the rank is deficient, and conduct inference on the coefficients on the alleged cointegrating vector using standard inference. Our theory shows that the inference is then severely distorted. When the rank is deficient or nearly deficient it is incorrect to use standard inference on the cointegrating vectors. Nonetheless, applying standard inference in the particular example leads to marginal rejection of the hypothesis. Applying the bounds test of Khalaf and Urga (2014) shifts the distribution to the right and there is not much power to reject a hypothesis. If the rank is deficient, which is possible in the example, the alleged cointegrating vector cannot be cointegrating.

Rank deficiency also matters when the rank is determined empirically. Different asymptotic distributions arise in the standard case and when the rank is deficient. The asymptotic distribution tends to give a very good approximation to the finite sample distribution when the rank is far from being deficient, see for instance Nielsen $(1997,2004)$ When the parameters are in the vicinity of rank deficiency the finite sample distribution tends to be a combination of the two asymptotic distributions. When the parameters are not too close to the rank deficient case a Bartlett correction using a fixed parameter second-order asymptotic expansion works very well, see Johansen $(2000,2002)$ Bootstrap solutions have been discussed in simulation studies by Fachin (2000); Gredenhoff and Jacobson (2001); Swensen (2004); Cavaliere et al. (2012). When the parameters are closer to rank deficient a local-to-unity asymptotic expansion gives an improvement, see Nielsen (2004) for the cointegration case and Nielsen $(1999,2001)$ for the corresponding instrumental variable case. A starting point for the finite sample analysis is knowledge of the fixed-parameter first-order asymptotic theory across the parameter space, including rank deficient cases.

We discuss the asymptotic theory for models without and with deterministic terms in Sections 2 and 3, respectively. The implications for finite sample analysis and the weakly identified case are discussed in Section 4 along with an application to US treasury zero coupon yields. Section 5 concludes. Proofs are given in an Appendix A.

## 2. The Model without Deterministic Terms

We consider the Gaussian cointegrated vector autoregressive model in the case with no deterministic terms. The asymptotic theory for tests for reduced cointegration rank and for a known cointegrating vector is derived when the rank is deficient. Finally, we analyze the case of near rank deficiency.

### 2.1. Model and Hypotheses

Consider a $p$-dimensional time series $X_{t}$ for $t=1-k, \ldots, 0,1, \ldots T$. The unrestricted vector autoregressive model can be written as

$$
\begin{equation*}
\Delta X_{t}=\Pi X_{t-1}+\sum_{i=1}^{k-1} \Gamma_{i} \Delta X_{t-i}+\varepsilon_{t} \quad \text { for } t=1, \ldots, T \tag{1}
\end{equation*}
$$

where the innovations $\varepsilon_{t}$ are independent normal $\mathrm{N}_{p}(0, \Omega)$-distributed. The parameters $\Pi, \Gamma_{i}, \Omega$ are freely varying $p$-dimensional square matrices so that $\Omega$ is symmetric, positive definite.

The hypothesis of reduced cointegration rank is formulated as

$$
\begin{equation*}
\mathrm{H}_{z}(r): \quad \operatorname{rank} \Pi \leq r \tag{2}
\end{equation*}
$$

for some $0 \leq r \leq p$. The interpretation of the hypotheses follows from the Granger-Johansen representation presented in Section 2.2 below. The subscript $z$ indicates that the model has a zero deterministic component. The rank hypotheses are nested so that

$$
\begin{equation*}
\mathrm{H}_{z}(0) \subset \cdots \subset \mathrm{H}_{z}(r) \subset \cdots \subset \mathrm{H}_{z}(p) \tag{3}
\end{equation*}
$$

The rank deficiency problem arises when testing the hypothesis $\mathrm{H}_{z}(r)$ when in fact the sub-hypothesis $\mathrm{H}_{z}(r-1)$ is satisfied. The rank is determined to be $r$ if the hypothesis $\mathrm{H}_{z}(r)$ cannot be rejected while the sub-hypothesis $\mathrm{H}_{z}(r-1)$ is rejected. As a short-hand we write $\mathrm{H}_{z}^{\circ}(r)=\mathrm{H}_{z}(r) \backslash \mathrm{H}_{z}(r-1)$ for this situation. The rank can be determined along the procedure outlined in Johansen $(1992,1995)$ [Section 12.1] and Paruolo (2001). In practice, these decisions are often marginal, hence the need to study the asymptotic theory of test statistics in the rank deficient case.

The rank hypothesis can equivalently be written as

$$
\begin{equation*}
\mathrm{H}_{z}(r): \quad \Pi=\alpha \beta^{\prime}, \tag{4}
\end{equation*}
$$

where $\alpha$ and $\beta$ are $p \times r$ matrices. The advantage of this formulation is that $\alpha$ and $\beta$ vary in vector spaces. The formulation does, however, allow rank deficiency where the rank of $\Pi$ is smaller than $r$. We follow Johansen (1991, Equation (2.2)) and refer to $\beta$ as the cointegrating vectors. We find the terminology useful, although it is ambiguous. Indeed, for a particular data generating process where $\Pi$ has rank less than $r$ then the identity $\Pi=\alpha \beta^{\prime}$ can be satisfied while columns of $\beta$ may not be row-eigenvectors of $\Pi$ in which case $\beta^{\prime} X_{t}$ cannot be stationary. Even when $\Pi$ has rank $r$ then $\beta^{\prime} X_{t}$ is only (approximately) stationary under the $\mathrm{I}(1)$ condition introduced below. However, from a statistical viewpoint, the estimator of $\Pi$ under the restriction of rank $r$ will in a finite sample have rank $r$ with probability one. In practice our only knowledge of the rank arises from inference. Johansen's terminology appears to be focused on the statistical viewpoint which we will follow even when studying the rank deficient cases.

The hypothesis of known cointegration vectors is

$$
\begin{equation*}
\mathrm{H}_{z, \beta}(r): \quad \Pi=\alpha b^{\prime} \tag{5}
\end{equation*}
$$

for some unknown matrix $\alpha$ and a known matrix $b$, both of dimension $p \times r$, so that $b$ has full column rank. The standard analysis is concerned with the situation where $\alpha$ has full column rank, but in the rank deficient case, it has reduced column rank, so that the hypothesis $\mathrm{H}_{z}(r-1)$ is satisfied. When referring to $b$ as the cointegrating vectors, we, once again, follow the terminology of Johansen (1991, Equation (3.1)) even though $b^{\prime} X_{t}$ cannot be stationary under rank deficiency.

### 2.2. Granger-Johansen Representation

The Granger-Johansen representation provides an interpretation of the cointegration model that is useful in the asymptotic analysis. We work with the result stated by Johansen (1995, Theorem 4.2). The theorem requires the following assumption.
$\mathbf{I}(1)$ Condition. Suppose $\operatorname{rank} \Pi=s$ where $s \leq p$. Consider the characteristic roots satisfying $0=\operatorname{det}\{A(z)\}$ where $A(z)=(1-z) I_{p}-\Pi z-\sum_{i=1}^{k-1} \Gamma_{i} z^{i}(1-z)$. Suppose there are $p-s$ unit roots, and that the remaining roots are stationary roots, so satisfying $|z|>1$.

The Granger-Johansen theorem assumes that a process satisfying the model (1) so that rank $\Pi=$ $r$ and we can write $\Pi=\alpha \beta^{\prime}$ while the $\mathrm{I}(1)$ condition holds with $s=r$. The process then has the representation

$$
\begin{equation*}
X_{t}=C \sum_{i=1}^{t} \varepsilon_{i}+S_{t}+\tau \tag{6}
\end{equation*}
$$

where the impact matrix $C$ for the random walk has rank $p-r$ and satisfies $\beta^{\prime} C=0$ and $C \alpha=0$, the process $S_{t}$ can be given a zero mean stationary initial distribution and $\tau$ depends on the initial observations in such a way that $\beta^{\prime} \tau=0$. In other words, the process $X_{t}$ behaves like a random walk with cointegrating relations $\beta^{\prime} X_{t}$ that can be given a stationary initial distribution.

### 2.3. Test Statistics

The likelihood ratio test statistic for the reduced rank hypothesis $\mathrm{H}_{z}(r)$ against the unrestricted model $\mathrm{H}_{z}(p)$ is found by reduced rank regression, see Johansen (1995, Section 6). It can be described as a two-step procedure. First, the differences $\Delta X_{t}$ and the lagged levels $X_{t-1}$ are regressed on the lagged differences $\Delta X_{t-i}, i=1, \ldots, k-1$ giving residuals $R_{0, t}, R_{1, t}$. Secondly, the squared sample correlations, $1 \geq \widehat{\lambda}_{1} \geq \cdots \geq \widehat{\lambda}_{p} \geq 0$ say, of $R_{0, t}$ and $R_{1, t}$ are found, by computing product moments $S_{i j}=T^{-1} \sum_{t=1}^{T} R_{i, t} R_{j, t}^{\prime}$ and solving the eigenvalue problem $0=\operatorname{det}\left(\lambda S_{11}-S_{10} S_{00}^{-1} S_{01}\right)$. The log likelihood ratio test statistic for the rank hypothesis is then

$$
\begin{equation*}
L R\left\{\mathrm{H}_{z}(r) \mid \mathrm{H}_{z}(p)\right\}=-T \sum_{j=r+1}^{p} \log \left(1-\widehat{\lambda}_{j}\right) \tag{7}
\end{equation*}
$$

Under the hypothesis of known cointegration vectors, the likelihood is maximised by least squares regression. The log likelihood ratio test statistic against the unrestricted model $\mathrm{H}_{z}(p)$ is therefore given by

$$
\begin{equation*}
L R\left\{\mathrm{H}_{z, \beta}(r) \mid \mathrm{H}_{z}(p)\right\}=-T \log \frac{\operatorname{det}\left\{S_{00}-S_{01} S_{11}^{-1} S_{10}\right\}}{\operatorname{det}\left\{S_{00}-S_{01} b\left(b^{\prime} S_{11} b\right)^{-1} b^{\prime} S_{10}\right\}} . \tag{8}
\end{equation*}
$$

The log likelihood ratio statistic for the hypothesis of known cointegrating vector against the rank hypothesis is found by combining the statistics in (7) and (8), that is

$$
\begin{equation*}
L R\left\{\mathrm{H}_{z, \beta}(r) \mid \mathrm{H}_{z}(r)\right\}=\operatorname{LR}\left\{\mathrm{H}_{z, \beta}(r) \mid \mathrm{H}_{z}(p)\right\}-\operatorname{LR}\left\{\mathrm{H}_{z}(r) \mid \mathrm{H}_{z}(p)\right\} . \tag{9}
\end{equation*}
$$

The relationship will be useful in the asymptotic theory. For instance, Theorems 1 and 2 give the asymptotic distributions of $L R\left\{\mathrm{H}_{z}(r) \mid \mathrm{H}_{z}(p)\right\}$ and $L R\left\{\mathrm{H}_{z, \beta}(r) \mid \mathrm{H}_{z}(p)\right\}$, respectively. From this we can derive an expression for the distribution of $\operatorname{LR}\left\{\mathrm{H}_{z, \beta}(r) \mid \mathrm{H}_{z}(r)\right\}$. When it comes to tabulation we will need to simulate all three distributions. This would be the case even if the former two statistics were independent.

### 2.4. Asymptotic Theory for the Rank Test

In the asymptotic analysis it is possible to relax the assumption to the innovations. While the likelihood is derived under the assumption of independent, identically Gaussian distributed innovations less is needed for the asymptotic theory. Johansen (1995) assumes the innovations are independent, identically distributed with mean zero and finite variance and uses linear process results from Phillips and Solo (1992). This could be relaxed further to, for instance, a martingale difference assumption. However, for expositional simplicity we follow Johansen's argument and assumptions.

Theorem 1. Consider the rank hypothesis $\mathrm{H}_{z}(r): \operatorname{rank} \Pi \leq r$. Suppose $\mathrm{H}_{z}^{\circ}(s)=\mathrm{H}_{z}(s) \backslash \mathrm{H}_{z}(s-1)$ holds for some $s \leq r$ and that the $I(1)$ condition holds for that s. Let $F_{u}=B_{u}$ be a $p$-s-dimensional standard Brownian motion on $[0,1]$. Let $1 \geq \rho_{1} \geq \cdots \geq \rho_{p-s} \geq 0$ be the eigenvalues of the eigenvalue problem

$$
\begin{equation*}
0=\operatorname{det}\left\{\rho \int_{0}^{1} F_{u} F_{u}^{\prime} d u-\int_{0}^{1} F_{u}\left(d B_{u}\right)^{\prime} \int_{0}^{1}\left(d B_{u}\right) F_{u}^{\prime}\right\} \tag{10}
\end{equation*}
$$

Then, for $T \rightarrow \infty$,

$$
\begin{equation*}
L R\left\{\mathrm{H}_{z}(r) \mid \mathrm{H}_{z}(p)\right\}=-T \sum_{j=r+1}^{p} \log \left(1-\widehat{\lambda}_{j}\right) \xrightarrow{\mathrm{D}} \sum_{j=r-s+1}^{p-s} \rho_{j} . \tag{11}
\end{equation*}
$$

In the standard non-deficient situation where $r=s$ the result reduces to the result of Johansen (1995, Theorem 6.1). The rank deficient case was also discussed by Johansen (1995, p. 158) and Nielsen (2004, Theorem 6.1).

Table 1 reports the asymptotic distribution of the rank test reported in Theorem 1. The simulation were done using Ox (Doornik 2007). The simulation design follows that of Johansen (1995, Section 15). That is, the stochastic integrals in (10) were descretized with $T=1000$ and zero initial observations with one million repetitions. The table reports simulated quantiles and moments for $r-s=0,1,2$ and $p-r=1,2,3,4$. However, the case of $p-r=1$ and $r-s=0$ are analytic values from Nielsen (1997) and where the quantiles were provided by Karim Abadir using his results in Abadir (1995). Bernstein (2014) reports values for higher dimensions. The $85 \%$ quantile has not been computed analytically in this case.

Table 1. Quantiles, mean, and variance of $L R\left\{\mathrm{H}_{z}(r) \mid \mathrm{H}_{z}(p)\right\}$, where the data generating process has $\operatorname{rank} s=\operatorname{rank} \Pi \leq r$.

| $r-s$ | $p-r$ | $\mathbf{5 0 \%}$ | $\mathbf{8 0 \%}$ | $\mathbf{8 5 \%}$ | $\mathbf{9 0 \%}$ | $\mathbf{9 5 \%}$ | $\mathbf{9 7 . 5 \%}$ | $\mathbf{9 9 \%}$ | Mean | Var |
| :--- | :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 0 | 1 | 0.60 | 1.88 | - | 2.98 | 4.13 | 5.32 | 6.94 | 1.14 | 2.22 |
|  | 2 | 5.48 | 8.48 | 9.31 | 10.44 | 12.30 | 14.07 | 16.34 | 6.09 | 10.61 |
|  | 3 | 14.39 | 18.94 | 20.13 | 21.70 | 24.22 | 26.54 | 29.37 | 15.02 | 25.13 |
|  | 4 | 27.29 | 33.35 | 34.88 | 36.91 | 40.04 | 42.93 | 46.45 | 27.93 | 45.66 |
| 1 | 1 | 0.36 | 1.13 | 1.38 | 1.74 | 2.35 | 2.98 | 3.81 | 0.67 | 0.70 |
|  | 2 | 4.27 | 6.25 | 6.78 | 7.50 | 8.65 | 9.76 | 11.14 | 4.61 | 4.66 |
|  | 3 | 11.92 | 15.20 | 16.04 | 17.14 | 18.88 | 20.50 | 22.48 | 12.31 | 13.22 |
|  | 4 | 23.47 | 28.09 | 29.25 | 30.76 | 33.10 | 35.21 | 37.83 | 23.89 | 26.96 |
| 2 | 1 | 0.30 | 0.97 | 1.18 | 1.48 | 1.98 | 2.47 | 3.11 | 0.56 | 0.48 |
|  | 2 | 3.93 | 5.57 | 6.01 | 6.59 | 7.51 | 8.38 | 9.46 | 4.18 | 3.24 |
|  | 3 | 11.04 | 13.82 | 14.53 | 15.46 | 16.91 | 18.24 | 19.87 | 11.34 | 9.63 |
|  | 4 | 21.84 | 25.83 | 26.82 | 28.11 | 30.09 | 31.91 | 34.13 | 22.18 | 20.21 |

The first panel of Table 1 reports the distribution for the standard case where $s=r$. This corresponds to Table 15.1 of Johansen (1995). The second and third panel of Table 1 report the distribution for the rank deficient case where $s=r-1$ so $r-s=1$ and where $s=r-2$ so $r-s=2$. The first entry in panel 2 for $s=r-1$ and $p-r=1$, so $r-s=1$, corresponds to Table 6 of Nielsen (2004). It is seen that as the rank becomes more deficient the distribution shifts to the left. It should be noted that if the rank is non deficient, but the $\mathrm{I}(1)$ condition is not satisfied then the distribution would tend to shift to the right, see Nielsen (2004) for a discussion. The simulations reported in Table 8 of that paper indicates that the distribution is between these extremes if the rank is deficient and the $\mathrm{I}(1)$ condition fails.

The rank test statistic in (7) has been analyzed analytically for the canonical correlation problem in cross-sectional models in Nielsen $(1999,2001)$ This test also corresponds to the test for relevance in the instrument variable problem. In that case, analytic expressions are available when $p=2, r=1$
and $s=0,1$. When $s=1$ we have a $\chi^{2}$-distribution with mean 1 and variance 2 . When $s=0$ the mean is 0.429 and the variance is $0.575-(0.429)^{2}=0.391$, see Nielsen (1999). Thus, the impact of rank deficiency is similar to what is seen in Table 1 for cointegration rank testing.

### 2.5. Asymptotic Theory for the Test on the Cointegrating Vectors

In the analysis of the test for known cointegrating vectors, we focus on the situation where the data generating process has rank $s=0$. In this situation the asymptotic distribution is relatively simple to describe, because it does not depend on the value of the hypothesized cointegrating vectors $b$. This is adequate for a discussion of aspects of situations considered in Khalaf and Urga (2014). If the rank is non-zero but deficient so $0<s<r$, then the data generating process will have cointegrating vectors $\beta_{0}$ of dimension $p \times s$ and the asymptotic theory will depend on $\beta_{0}$ and $b$. In practice, it is rare to test for simple hypotheses when there is more than one hypothesized cointegrating vector, so we do not pursue this complication.

The analysis of the test for known cointegrating vectors is somewhat different from the analysis in Johansen (1995). His analysis is aimed at the situation where different restrictions are imposed on the cointegrating vectors. The argument then involves an intriguing consistency proof for the estimated cointegrating vectors. However, when testing the hypothesis of known cointegrating vectors the likelihood is maximized by the least squares method and the consistency argument is not needed. The asymptotic theory can then be described by the following result.

Theorem 2. Consider the hypothesis $\mathrm{H}_{z, \beta}(r): \Pi=\alpha b^{\prime}$, where $\alpha, b$ have dimension $p \times r$ and where $\alpha$ is unknown and $b$ is known with full column rank. Suppose $\mathrm{H}_{z}(0)$ is satisfied, so that $\alpha=0$ and $s=0$, and that the $I(1)$ condition is satisfied with $s=0$. Let $B_{u}$ be a $p$-dimensional standard Brownian motion on $[0,1]$ with components $B_{1, u}$ and $B_{2, u}$ of dimension $r$ and $p-r$, respectively. Then, for $T \rightarrow \infty$,

$$
\begin{align*}
& L R\left\{\mathrm{H}_{z, \beta}(r) \mid \mathrm{H}_{z}(p)\right\} \stackrel{\mathrm{D}}{\rightarrow} \operatorname{tr}\left\{\int_{0}^{1} d B_{u} B_{u}^{\prime}\left(\int_{0}^{1} B_{u} B_{u}^{\prime} d u\right)^{-1} \int_{0}^{1} B_{u}\left(d B_{u}\right)^{\prime}\right. \\
&\left.\quad-\int_{0}^{1} d B_{u} B_{1, u}^{\prime}\left(\int_{0}^{1} B_{1, u} B_{1, u}^{\prime} d u\right)^{-1} \int_{0}^{1} B_{1, u}\left(d B_{u}\right)^{\prime}\right\} . \tag{12}
\end{align*}
$$

The convergence of the test statistic $\operatorname{LR}\left\{\mathrm{H}_{z, \beta}(r) \mid \mathrm{H}_{z}(p)\right\}$ holds jointly with the convergence for the rank test statistic $\operatorname{LR}\left\{\mathrm{H}_{z}(r) \mid \mathrm{H}_{z}(p)\right\}$, for $s=0$, in Theorem 1. Thus, when $s=0$ the formula (9) implies that the limit distribution of the test statistic for known $\beta$ within the model with rank of at most $r$ can be found as the difference of the two limiting variables.

Table 2 reports the asymptotic distribution of the test for known cointegrating vector in the model where the rank is at most $r$. When $s=r$ the asymptotic distribution is $\chi^{2}$ with $r(p-r)$ degrees of freedom, see Johansen (1995, Theorem 7.2.1). When $s=0$ the asymptotic distribution reported in Theorem 2 applies. The simulation design is as before. It is seen that in the rank deficient case the distribution is shifted to the right. This matches the finite sample simulations reported by Johansen (2000, Table 2).

Table 2. Quantiles, mean, and variance of $\operatorname{LR}\left\{\mathrm{H}_{z, \beta}(r) \mid \mathrm{H}_{z}(r)\right\}$, where the data generating process has $\operatorname{rank} s=\operatorname{rank} \Pi \leq r$.

| $p$ | $r$ | $\boldsymbol{s}$ | $\mathbf{5 0 \%}$ | $\mathbf{8 0 \%}$ | $\mathbf{8 5 \%}$ | $\mathbf{9 0 \%}$ | $\mathbf{9 5 \%}$ | $\mathbf{9 7 . 5 \%}$ | $\mathbf{9 9 \%}$ | Mean | Var |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 2 | 1 | 1 | 0.45 | 1.64 | 2.07 | 2.71 | 3.84 | 5.02 | 6.63 | 1 | 2 |
|  |  | 0 | 2.62 | 5.44 | 6.22 | 7.30 | 9.05 | 10.75 | 12.96 | 3.31 | 8.71 |
| 3 | 2 | 2 | 1.39 | 3.22 | 3.79 | 4.61 | 5.99 | 7.38 | 9.21 | 2 | 4 |
|  |  | 0 | 5.80 | 9.42 | 10.40 | 11.71 | 13.82 | 15.77 | 18.27 | 6.42 | 15.53 |
| 3 | 1 | 1 | 1.39 | 3.22 | 3.79 | 4.61 | 5.99 | 7.38 | 9.21 | 2 | 4 |
|  |  | 0 | 6.79 | 10.58 | 11.57 | 12.89 | 15.02 | 17.00 | 19.49 | 7.33 | 17.52 |

Table 3 reports the simulated asymptotic distribution of the test for known cointegrating vector in the model where the rank is unrestricted. The distribution is shifted to the right in the rank deficient case. Note, that the table reports the distribution of the convolution of the statistics simulated in Tables 1 and 2, see (9). Thus, up to a simulation error the expectations reported in Tables 1 and 2 add up to the expectation reported in Table 3. In the full rank case $r=s$ the statistics in Tables 1 and 2 are independent, as proved below, so also the variances are additive.

Theorem 3. Consider the hypothesis $\mathrm{H}_{z, \beta}^{\circ}(r)$. Suppose $\mathrm{H}_{z}^{\circ}(r)=\mathrm{H}_{z}^{\circ}(r) / \mathrm{H}_{z}^{\circ}(r-1)$ is satisfied and that the $I(1)$ condition holds with $s=r$. Then the rank test statistic $\operatorname{LR}\left\{\mathrm{H}_{z}^{\circ}(r) \mid \mathrm{H}_{z}^{\circ}(p)\right\}$ and the statistic $\operatorname{LR}\left\{\mathrm{H}_{z, \beta}^{\circ}(r) \mid \mathrm{H}_{z}^{\circ}(r)\right\}$ for testing a simple hypothesis on the cointegrating vector are asymptotically independent.

The asymptotic distribution of the rank statistic $\operatorname{LR}\left\{\mathrm{H}_{z}^{\circ}(r) \mid \mathrm{H}_{z}^{\circ}(p)\right\}$ is given in Theorem 1, while the statistic for the cointegrating vector $\operatorname{LR}\left\{\mathrm{H}_{z, \beta}^{\circ}(r) \mid \mathrm{H}_{z}^{\circ}(r)\right\}$ is asymptotically $\chi^{2}\{r(p-r)\}$.

Table 3. Quantiles, mean, and variance of $L R\left\{\mathrm{H}_{z, \beta}(r) \mid \mathrm{H}_{z}(p)\right\}$, where the data generating process has $\operatorname{rank} s=\operatorname{rank} \Pi \leq r$.

| $\boldsymbol{p}$ | $\boldsymbol{r}$ | $\boldsymbol{s}$ | $\mathbf{5 0 \%}$ | $\mathbf{8 0 \%}$ | $\mathbf{8 5 \%}$ | $\mathbf{9 0 \%}$ | $\mathbf{9 5 \%}$ | $\mathbf{9 7 . 5} \%$ | $\mathbf{9 9 \%}$ | Mean | Var |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 2 | 1 | 1 | 1.54 | 3.43 | 4.01 | 4.83 | 6.22 | 7.62 | 9.47 | 2.15 | 4.23 |
|  |  | 0 | 3.35 | 6.11 | 6.89 | 7.95 | 9.70 | 11.38 | 13.57 | 3.98 | 8.82 |
| 3 | 2 | 2 | 2.52 | 4.85 | 5.53 | 6.48 | 8.07 | 9.60 | 11.62 | 3.15 | 6.26 |
|  |  | 0 | 6.36 | 9.96 | 10.92 | 12.22 | 14.32 | 16.29 | 18.79 | 6.98 | 15.35 |
| 3 | 1 | 1 | 7.50 | 11.03 | 11.98 | 13.27 | 15.34 | 17.30 | 19.81 | 8.13 | 14.73 |
|  |  | 0 | 11.33 | 15.73 | 16.88 | 18.41 | 20.83 | 23.09 | 25.91 | 11.96 | 23.31 |

### 2.6. The Case of Nearly Deficient Rank

With the above results we have two extremes. First, the full rank case where standard results apply, that is Johansen's Dickey-Fuller type distribution for rank testing and $\chi^{2}$ inferences for testing constraints on the cointegrating vectors. Second, the rank deficient case where new Dickey-Fuller type distributions apply both for rank testing and for testing constraints on the cointegrating vectors. In between these extremes we have the nearly rank deficient case corresponding to weak identification in the instrumental variable literature. These nearly deficient cases can be analyzed using local-to-unity parametrization. However, a full theory is notationally complicated as there will be many nuisance parameters. We therefore consider a simple special case inspired by the power analysis of Johansen (1995, Section 14) and distribution analysis of Nielsen (2004).

The main finding is that the appropriate local rate is $T^{-1}$ as in power analysis for unit tests and cointegration rank tests as opposed to $T^{-1 / 2}$ for stationary models as in Andrews and Cheng (2012). Consider a bivariate, first order, local-to-unity vector autoregressive model where

$$
\Delta X_{t}=\frac{1}{T}\left(\begin{array}{cc}
b_{1} & b_{2}  \tag{13}\\
0 & 0
\end{array}\right) X_{t-1}+\varepsilon_{t} \quad \text { for } t=1, \ldots, T
$$

where the innovations $\varepsilon_{t}$ are independent normal $\mathrm{N}_{2}\left(0, I_{2}\right)$-distributed where $b_{1} \neq 0$.
We now have the following variant of the result for the rank test in Theorem 1.
Theorem 4 (Nielsen 2004, Theorem 6.2). Consider the data generating process (13). Let $B_{u}$ be a bivariate standard Brownian motion on $[0,1]$ and let $J_{u}$ be the bivariate Ornstein-Uhlenbeck process given by

$$
J_{u}=\left(\begin{array}{cc}
b_{1} & b_{2} \\
0 & 0
\end{array}\right) \int_{0}^{u} J_{s} d s+B_{u} .
$$

Let $1 \leq \rho_{1} \leq \rho_{2} \leq 0$ be the eigenvalues of the eigenvalue problem

$$
0=\operatorname{det}\left\{\rho \int_{0}^{1} J_{u} J_{u}^{\prime} d u-\int_{0}^{1} J_{u}\left(d B_{u}\right)^{\prime} \int_{0}^{1}\left(d B_{u}\right) J_{u}^{\prime}\right\}
$$

Then, for $T \rightarrow \infty$,

$$
L R\left\{\mathrm{H}_{z}(1) \mid \mathrm{H}_{z}(2)\right\} \xrightarrow{\mathrm{D}} \rho_{2}
$$

The limit distribution is tabulated in Nielsen (2004, Table 8).
We now consider the test for known cointegrating vector, $b=\left(b_{1}, b_{2}\right)^{\prime}$. The result in Theorem 2 is modified as follows.

Theorem 5. Consider the data generating process (13). Let $B_{u}, J_{u}$ be defined as in Theorem 4 and let $J_{1, u}=$ $b^{\prime} J_{u}$. Then

$$
\begin{aligned}
& L R\left\{\mathrm{H}_{z, \beta}(1) \mid \mathrm{H}_{z}(2)\right\} \stackrel{\mathrm{D}}{\rightarrow} \operatorname{tr}\left\{\int_{0}^{1} d B_{u} J_{u}^{\prime}\left(\int_{0}^{1} J_{u} J_{u}^{\prime} d u\right)^{-1} \int_{0}^{1} J_{u}\left(d B_{u}\right)^{\prime}\right. \\
&\left.\int_{0}^{1} d B_{u} J_{1, u}^{\prime}\left(\int_{0}^{1} J_{1, u} J_{1, u}^{\prime} d u\right)^{-1} \int_{0}^{1} J_{1, u}\left(d B_{u}\right)^{\prime}\right\}
\end{aligned}
$$

## 3. The Model with a Constant

We now consider the model augmented with a constant. In the cointegrated model the constant is restricted to the cointegrating space. Thus, the cointegrating vectors consist of vectors relating the dynamic variable extended by a further coordinate for the constant. There are now two rank conditions; one related to the dynamic part of these extended cointegrating vectors and one relating to the deterministic part of the cointegrating vectors. The condition to the cointegration rank in the standard theory can therefore fail in two ways.

### 3.1. Model and Hypotheses

The unrestricted vector autoregressive model is

$$
\begin{equation*}
\Delta X_{t}=\Pi X_{t-1}+\mu+\sum_{i=1}^{k-1} \Gamma_{i} \Delta X_{t-i}+\varepsilon_{t} \quad \text { for } t=1, \ldots, T \tag{14}
\end{equation*}
$$

where the innovations $\varepsilon_{t}$ are independent normal $\mathrm{N}_{p}(0, \Omega)$-distributed. The parameters are the $p$-dimensional square matrices $\Pi, \Gamma_{i}, \Omega$ and the $p$-vector $\mu$. They vary freely so that $\Omega$ is symmetric, positive definite.

For the model with a constant there are two types of cointegration rank hypotheses:

$$
\begin{align*}
\mathrm{H}_{c l}(r): & \operatorname{rank} \Pi & \leq r,  \tag{15}\\
\mathrm{H}_{c}(r): & \operatorname{rank}(\Pi, \mu) & \leq r . \tag{16}
\end{align*}
$$

Their interpretations follow from the Granger-Johansen representation which is reviewed in Section 3.2 below. In short, if there are no rank deficiencies the first hypothesis $\mathrm{H}_{c \ell}$ gives cointegrating relations with a constant level and common trends with a linear trend. The second hypothesis $\mathrm{H}_{c}$ has a constant level both for the cointegrating relations and the common trends. The hypotheses are nested so that

$$
\begin{equation*}
\mathrm{H}_{c}(0) \subset \mathrm{H}_{c \ell}(0) \subset \cdots \subset \mathrm{H}_{c \ell}(r-1) \subset \mathrm{H}_{c}(r) \subset \mathrm{H}_{c \ell}(r) \subset \cdots \subset \mathrm{H}_{c}(p)=\mathrm{H}_{c \ell}(p) . \tag{17}
\end{equation*}
$$

This nesting structure is considerably more complicated than the structure (3) for the model without deterministic terms. A practical investigation may start in three different ways. First, the model (14) is taken as the starting point. Both types of hypotheses come into play and the rank is determined as outlined in Johansen (1995, Section 12). Secondly, if visual inspection of the data indicates that linear trends are not present the hypotheses $\mathrm{H}_{c \ell}$ may be ignored. Thirdly, if visual inspection of the data indicates that a linear trend could be present, the model (14) should be augmented with a linear trend term and we move outside the present framework. Nielsen and Rahbek (2000) discuss the latter two possibilities. Here, we are concerned with the first two possibilities.

The rank hypotheses can equivalently be formulated as

$$
\begin{array}{rlrl}
\mathrm{H}_{c \ell}(r) & : & \Pi & =\alpha \beta^{\prime} \\
\mathrm{H}_{c}(r): & (\Pi, \mu) & =\alpha\left(\beta^{\prime}, \beta_{c}^{\prime}\right) . \tag{19}
\end{array}
$$

The hypotheses of known cointegrating vectors are therefore

$$
\begin{array}{rlrl}
\mathrm{H}_{c l, \beta}(r) & : & \Pi & =\alpha b^{\prime}, \\
\mathrm{H}_{c, \beta}(r): & (\Pi, \mu) & =\alpha\left(b^{\prime}, b_{c}^{\prime}\right) . \tag{21}
\end{array}
$$

for a known $(p \times r)$-matrix $b$ with full column rank and, in the second case, also a known $(1 \times r)$-matrix $b_{c}$ so that $b^{*}=\left(b^{\prime}, b_{c}^{\prime}\right)^{\prime}$ has full column rank.

### 3.2. Granger-Johansen Representation

We give a Granger-Johansen representation for each of the two reduced rank hypotheses. Both results follow from Theorem 4.2 and Exercise 4.5 of Johansen (1995). First, consider the hypothesis $\mathrm{H}_{c \ell}(r)$. Suppose that the sub-hypothesis $\mathrm{H}_{c}(r)$ does not hold and that the $\mathrm{I}(1)$ condition holds with $s=r$. Thus, the $(p \times r)$-matrices $\alpha, \beta$ have full column rank but $\alpha_{\perp}^{\prime} \mu \neq 0$, so that the matrix $\Pi^{*}=(\Pi, \mu)$ has rank $r+1$. Then, the Granger-Johansen representation is

$$
\begin{equation*}
X_{t}=C \sum_{i=1}^{t} \varepsilon_{i}+S_{t}+\tau_{c}+\tau_{\ell} t \tag{22}
\end{equation*}
$$

where the impact matrix $C$ has rank $p-r$ and satisfies $\beta^{\prime} C=0$ and $C \alpha=0$ while $\tau_{\ell}=C \mu \neq 0$. As a consequence, the process has a linear trend, but the cointegrating relations $\beta^{\prime} X_{t}$ do not have a linear trend, since $\beta^{\prime} C=0$.

Secondly, consider the hypothesis $\mathrm{H}_{c}(r)$. Suppose that the sub-hypothesis $\mathrm{H}_{c \ell}(r-1)$ does not hold and that the $\mathrm{I}(1)$ condition holds with $s=r$. Thus, the $(p \times r)$-matrices $\alpha, \beta$ have full column rank, and the $\{(p+1) \times r\}$-matrix $\beta^{*}=\left(\beta, \beta_{c}^{\prime}\right)^{\prime}$ has full column rank. Then, the Granger-Johansen representation (22) holds with $\tau_{\ell}=0$, while $\tau_{c}$ has the property that $\beta^{\prime} \tau_{c}=-\beta_{c}^{\prime}$. In other words, the process $X_{t}$ behaves like a random walk where $\beta^{\prime} X_{t}$ has an invariant distribution with a non-zero mean, while $\beta^{\prime} X_{t}+\beta_{c}^{\prime}$ has a zero mean invariant distribution.

### 3.3. Test Statistics

The test statistics are variations of those for the model without deterministic terms. The differences relate to the formation of the residuals $R_{0, t}$ and $R_{1, t}$

First, consider the reduced rank hypothesis $\mathrm{H}_{c \ell}(r)$ and the corresponding hypothesis $\mathrm{H}_{c \ell, \beta}(r)$ of known cointegrating vectors. The residuals $R_{0, t}$ and $R_{1, t}$ are formed by regressing the differences $\Delta X_{t}$ and the lagged levels $X_{t-1}$ on an intercept and the lagged differences $\Delta X_{t-i}, i=1, \ldots, k-1$. In the second step, compute the canonical correlations $1 \geq \widehat{\lambda}_{1} \geq \cdots \geq \widehat{\lambda}_{p} \geq 0$ of $R_{0, t}$ and $R_{1, t}$. The rank test statistic $L R\left\{\mathrm{H}_{c \ell}(r) \mid \mathrm{H}_{c \ell}(p)\right\}$ then has the form (7). The test statistic for known cointegrating vectors $L R\left\{\mathrm{H}_{c \ell, \beta}(r) \mid \mathrm{H}_{c \ell}(p)\right\}$ has the form (8), using the same residuals $R_{0, t}$ and $R_{1, t}$, and the hypothesized cointegrating vectors $b$.

Secondly, consider the reduced rank hypothesis $\mathrm{H}_{c}(r)$ and the corresponding hypothesis $\mathrm{H}_{c, \beta}(r)$ of known cointegrating vectors. The residuals $R_{0, t}$ and $R_{1, t}$ are formed by regressing the differences $\Delta X_{t}$ and the vector formed by stacking the lagged levels and an intercept $X_{t-1}^{*}=\left(X_{t-1}^{\prime}, 1\right)^{\prime}$ on the lagged differences $\Delta X_{t-i}, i=1, \ldots, k-1$. In the second step, compute the canonical correlation of these $R_{0, t}$ and $R_{1, t}$. The rank test statistic $L R\left\{\mathrm{H}_{c}(r) \mid \mathrm{H}_{c}(p)\right\}$ then has the form (7). The test statistic for known cointegrating vectors $L R\left\{\mathrm{H}_{c, \beta}(r) \mid \mathrm{H}_{c}(p)\right\}$ has the form (8), using the same residuals $R_{0, t}$ and $R_{1, t}$, and the hypothesized cointegrating vectors $b^{*}=\left(b^{\prime}, b_{c}^{\prime}\right)^{\prime}$.

### 3.4. Asymptotic Theory for the Rank Tests

There are now four situations to consider. Indeed, the nesting structure in (17) shows that each of the two rank hypotheses $\mathrm{H}_{c \ell}(r)$ and $\mathrm{H}_{c}(r)$ can be rank deficient in two ways when either of $\mathrm{H}_{c \ell}^{\circ}(s)=\mathrm{H}_{c l}(s) / \mathrm{H}_{c}(s)$ or $\mathrm{H}_{c}^{\circ}(s)=\mathrm{H}_{c}(s) / \mathrm{H}_{c \ell}(s-1)$ holds. In three cases the limiting distribution is of the same form as in Theorem 1, albeit with a different limiting random function $F_{u}$. In the fourth case the limiting distribution has nuisance parameters. The nuisance parameter case arises when testing $\mathrm{H}_{c}(r)$ with a data generating process satisfying $\mathrm{H}_{c \ell}^{\circ}(s)=\mathrm{H}_{c \ell}(s) / \mathrm{H}_{c}(s)$. This is the case that can often be ruled out through visual inspection of the data as mentioned in Section 3.1.

We start with the test for the hypothesis $\mathrm{H}_{c l}(r)$ in the rank deficient case where $\mathrm{H}_{c \ell}^{\circ}(s)=\mathrm{H}_{c \ell}(s) / \mathrm{H}_{c}(s)$ holds for $s<r$. Johansen (1995) discusses the possibility $\mathrm{H}_{c}^{\circ}(r)$. The asymptotic theory is as follows.

Theorem 6. Consider the rank hypothesis $\mathrm{H}_{c \ell}(r): \operatorname{rank} \Pi \leq r$. Suppose $\mathrm{H}_{c \ell}^{\circ}(s)=\mathrm{H}_{c \ell}(s) \backslash \mathrm{H}_{c}(s)$ holds for some $s \leq r$, so that $\operatorname{rank} \Pi=s$ and $\operatorname{rank}(\Pi, \mu)=s+1$ and that the $I(1)$ condition is satisfied for that $s$. Let $B_{u}$ be a $(p-s)$-dimensional standard Brownian motion on $[0,1]$. Define a $(p-s)$-dimensional vector $F_{u}$ with coordinates

$$
F_{i, u}= \begin{cases}B_{i, u}-\bar{B}_{i} & \text { for } i=1, \ldots, p-s-1 \\ u-1 / 2 & \text { for } i=p-s\end{cases}
$$

Then $\operatorname{LR}\left\{\mathrm{H}_{c \ell}(r) \mid \mathrm{H}_{c \ell}(p)\right\}$ converges as in (11) using the present $F$.
Table 4 reports the simulated asymptotic distribution of the rank test reported in Theorem 6. The first panel gives the standard case where $s=r$ and corresponds to Johansen (1995, Table 15.3). For $p-r=1$ the asymptotic distribution is actually $\chi^{2}$ and the numbers are the standard numerically calculated ones rather than simulated ones. The second and the third panel report the distribution for the rank deficient case $\mathrm{H}_{c \ell}^{\circ}(s)$ where $\mathrm{H}_{c \ell}(s)$ holds, but $\mathrm{H}_{c}(s)$ fails. The distribution is shifted to the left when $r-s>0$ as in Table 1.

Table 4. Quantiles, mean, and variance of $L R\left\{\mathrm{H}_{c \ell}(r) \mid \mathrm{H}_{c \ell}(p)\right\}$, where the data generating process satisfies $\mathrm{H}_{c \ell}^{\circ}(s)=\mathrm{H}_{c \ell}(s) \backslash \mathrm{H}_{c}(s)$ with $s \leq r$.

| $r-s$ | $p-r$ | $\mathbf{5 0 \%}$ | $\mathbf{8 0 \%}$ | $\mathbf{8 5 \%}$ | $\mathbf{9 0 \%}$ | $\mathbf{9 5 \%}$ | $\mathbf{9 7 . 5 \%}$ | $\mathbf{9 9 \%}$ | Mean | Var |
| :--- | :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 0 | 1 | 0.45 | 1.64 | 2.07 | 2.71 | 3.84 | 5.02 | 6.63 | 1 | 2 |
|  | 2 | 7.61 | 11.09 | 12.04 | 13.30 | 15.35 | 17.27 | 19.74 | 8.24 | 14.29 |
|  | 3 | 18.66 | 23.72 | 25.03 | 26.76 | 29.47 | 31.95 | 34.99 | 19.29 | 31.38 |
|  | 4 | 33.52 | 40.07 | 41.71 | 43.86 | 47.22 | 50.21 | 53.94 | 34.15 | 53.86 |
| 1 | 1 | 0.38 | 1.33 | 1.66 | 2.13 | 2.93 | 3.72 | 4.74 | 0.79 | 1.08 |
|  | 2 | 6.01 | 8.34 | 8.96 | 9.78 | 11.10 | 12.34 | 13.87 | 6.37 | 6.53 |
|  | 3 | 15.49 | 19.14 | 20.08 | 21.30 | 23.21 | 24.99 | 27.14 | 15.88 | 16.73 |
|  | 4 | 28.82 | 33.82 | 35.07 | 36.70 | 39.20 | 41.50 | 44.27 | 29.24 | 31.96 |
| 2 | 1 | 0.34 | 1.19 | 1.47 | 1.87 | 2.55 | 3.19 | 4.00 | 0.69 | 0.79 |
|  | 2 | 5.43 | 7.34 | 7.84 | 8.51 | 9.57 | 10.56 | 11.81 | 5.70 | 4.46 |
|  | 3 | 14.17 | 17.26 | 18.04 | 19.05 | 20.64 | 22.09 | 23.86 | 14.48 | 12.00 |
|  | 4 | 26.62 | 30.92 | 31.98 | 33.38 | 35.52 | 37.46 | 39.79 | 26.95 | 23.82 |

The second case is the test for the same hypothesis $\mathrm{H}_{c \ell}(r)$ in the rank deficient case where $\mathrm{H}_{c}^{\circ}(s)=\mathrm{H}_{c}(s) / \mathrm{H}_{c \ell}(s-1)$ holds for $s \leq r$.

Theorem 7. Consider the rank hypothesis $\mathrm{H}_{c \ell}(r)$ : rank $\Pi \leq r$. Suppose $\mathrm{H}_{c}^{\circ}(s)=\mathrm{H}_{c}(s) \backslash \mathrm{H}_{c \ell}(s-1)$ holds for somes $\leq r$, so that rank $\Pi=\operatorname{rank} \Pi^{*}=s$ and that the $I(1)$ condition is satisfied for that s. Let $B_{u}$ be $a(p-s)$-dimensional standard Brownian motion on $[0,1]$. Define a $(p-s)$-dimensional vector $F_{u}$ as the de-meaned Brownian motion

$$
F_{u}=B_{u}-\bar{B}=B_{u}-\int_{0}^{1} B_{v} d v
$$

Then $\operatorname{LR}\left\{\mathrm{H}_{c \ell}(r) \mid \mathrm{H}_{c \ell}(p)\right\}$ converges as in (11) using the present $F$.
Table 5 reports the simulated asymptotic distribution of the rank test reported in Theorem 7. The first panel where $s=r$ and corresponds to Table A. 2 of Johansen and Juselius (1990). It is shifted to the right when compared to the first panel of Table 4. The second and the third panel of Table 5 report the distribution for the rank deficient case $\mathrm{H}_{c}^{\circ}(s)$ for $s<r$. In those case the distribution is shifted to the left relative to the first panel as in Tables 1 and 4.

Table 5. Quantiles, mean, and variance of $L R\left\{\mathrm{H}_{c \ell}(r) \mid \mathrm{H}_{c \ell}(p)\right\}$, where the data generating process satisfies $\mathrm{H}_{c}^{\circ}(s)=\mathrm{H}_{c}(s) \backslash \mathrm{H}_{c \ell}(s-1)$ with $s \leq r$.

| $r-s$ | $p-r$ | $\mathbf{5 0 \%}$ | $\mathbf{8 0 \%}$ | $\mathbf{8 5 \%}$ | $\mathbf{9 0 \%}$ | $\mathbf{9 5 \%}$ | $\mathbf{9 7 . 5 \%}$ | $\mathbf{9 9 \%}$ | Mean | Var |
| :--- | :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 0 | 1 | 2.45 | 4.90 | 5.60 | 6.56 | 8.15 | 9.72 | 11.71 | 3.04 | 6.95 |
|  | 2 | 9.39 | 13.36 | 14.41 | 15.80 | 18.03 | 20.14 | 22.80 | 10.03 | 18.66 |
|  | 3 | 20.30 | 25.70 | 27.09 | 28.89 | 31.75 | 34.37 | 37.61 | 20.95 | 35.73 |
|  | 4 | 35.19 | 42.01 | 43.71 | 45.94 | 49.38 | 52.52 | 56.31 | 35.84 | 58.26 |
| 1 | 1 | 1.51 | 3.12 | 3.55 | 4.12 | 5.04 | 5.92 | 7.03 | 1.87 | 2.72 |
|  | 2 | 7.21 | 9.95 | 10.66 | 11.61 | 13.09 | 14.47 | 16.21 | 7.60 | 8.95 |
|  | 3 | 16.78 | 20.75 | 21.75 | 23.08 | 25.13 | 26.98 | 29.32 | 17.20 | 19.57 |
|  | 4 | 30.25 | 35.49 | 36.81 | 38.51 | 41.15 | 43.56 | 46.46 | 30.69 | 35.22 |
| 2 | 1 | 1.16 | 2.54 | 2.89 | 3.36 | 4.09 | 4.76 | 5.62 | 1.48 | 1.81 |
|  | 2 | 6.38 | 8.66 | 9.25 | 10.03 | 11.26 | 12.40 | 13.80 | 6.69 | 6.23 |
|  | 3 | 15.27 | 18.64 | 19.49 | 20.61 | 22.35 | 23.94 | 25.88 | 15.61 | 14.27 |
|  | 4 | 28.00 | 32.45 | 33.58 | 35.05 | 37.32 | 39.37 | 41.85 | 28.26 | 26.55 |

In the third case we consider the test for the hypothesis $\mathrm{H}_{c}(r)$ in the rank deficient case where $\mathrm{H}_{c}^{\circ}(s)=\mathrm{H}_{c}(s) / \mathrm{H}_{c \ell}(s-1)$ holds for $s<r$.

Theorem 8. Consider the rank hypothesis $\mathrm{H}_{c}(r)$ : rank $\Pi \leq r$. Suppose $\mathrm{H}_{c}^{\circ}(s)=\mathrm{H}_{c}(s) \backslash \mathrm{H}_{c \ell}(s-1)$ holds for some $s \leq r$ so that rank $\Pi=\operatorname{rank}(\Pi, \mu)=s$ and that the $I(1)$ condition is satisfied for that s. Let $B_{u}$ be a $(p-s)$-dimensional standard Brownian motion on $[0,1]$. Define $a(p-s+1)$-dimensional vector $F_{u}$ given as

$$
\begin{equation*}
F_{u}=\binom{B_{u}}{1} \tag{23}
\end{equation*}
$$

Then $\operatorname{LR}\left\{\mathrm{H}_{c}(r) \mid \mathrm{H}_{c}(p)\right\}$ converges as in (11) using the present $F$.
Table 6 reports the simulated asymptotic distribution of the rank test reported in Theorem 8. The first panel gives the standard case where $s=r$ and corresponds to Johansen (1995, Table 15.2). The second and the third panel report the distribution for the rank deficient case $\mathrm{H}_{c}^{\circ}(s)$ for $s<r$. Once again, the distribution shifts to the left in the rank deficient case.

Table 6. Quantiles, mean, and variance of $\operatorname{LR}\left\{\mathrm{H}_{c}(r) \mid \mathrm{H}_{c}(p)\right\}$, where the data generating process satisfies $H_{c}^{\circ}(s)=H_{c}(s) \backslash H_{c \ell}(s-1)$ with $s \leq r$.

| $r-s$ | $p-r$ | $\mathbf{5 0 \%}$ | $\mathbf{8 0 \%}$ | $\mathbf{8 5 \%}$ | $\mathbf{9 0 \%}$ | $\mathbf{9 5 \%}$ | $\mathbf{9 7 . 5 \%}$ | $\mathbf{9 9 \%}$ | Mean | Var |
| :--- | :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 0 | 1 | 3.44 | 5.86 | 6.56 | 7.52 | 9.13 | 10.69 | 12.74 | 4.04 | 6.89 |
|  | 2 | 11.40 | 15.43 | 16.49 | 17.91 | 20.18 | 22.33 | 25.03 | 12.02 | 19.50 |
|  | 3 | 23.31 | 28.86 | 30.28 | 32.15 | 35.06 | 37.74 | 41.04 | 23.95 | 38.13 |
|  | 4 | 39.20 | 46.23 | 47.99 | 50.28 | 53.82 | 57.05 | 61.01 | 39.84 | 62.48 |
| 1 | 1 | 2.74 | 4.27 | 4.70 | 5.27 | 6.21 | 7.10 | 8.25 | 3.05 | 2.75 |
|  | 2 | 9.47 | 12.30 | 13.04 | 14.01 | 15.54 | 16.96 | 18.74 | 9.84 | 9.81 |
|  | 3 | 20.04 | 24.19 | 25.25 | 26.63 | 28.76 | 30.71 | 33.13 | 20.45 | 21.78 |
|  | 4 | 34.51 | 40.03 | 41.40 | 43.17 | 45.93 | 48.43 | 51.41 | 34.95 | 39.09 |
| 2 | 1 | 2.62 | 3.89 | 4.22 | 4.68 | 5.41 | 6.10 | 6.96 | 2.84 | 1.87 |
|  | 2 | 8.86 | 11.26 | 11.87 | 12.67 | 13.93 | 15.10 | 16.54 | 9.14 | 7.06 |
|  | 3 | 18.77 | 22.37 | 23.27 | 24.43 | 26.23 | 27.88 | 29.91 | 19.09 | 16.34 |
|  | 4 | 32.40 | 37.23 | 38.43 | 39.98 | 42.35 | 44.52 | 47.08 | 32.76 | 30.09 |

The final case is the test for the hypothesis $\mathrm{H}_{c}(r)$ in the rank deficient case where $\mathrm{H}_{c \ell}^{\circ}(s)=$ $\mathrm{H}_{c \ell}(s-1) / \mathrm{H}_{c}(s-1)$ for $s<r$. In this case the limiting distribution has nuisance parameters. We do not give the result here, since it is complicated to state and it does not seem particularly useful in practice. Indeed in practical work, this type of data generating process can often be ruled through visual data inspection as discussed in Section 3.1. Furthermore, it would be hard to deal with the nuisance parameters in applications.

It is worth noting that the proof in this final case would be somewhat different from the proof of Theorems 1, 6-8. They are all proved by modifying the argument of Johansen (1995, Sections 10 and 11). However, in the final case, a cointegration vector with random coefficients arise. Therefore, the analysis is best carried out in terms of the dual eigenvalue problem $0=\operatorname{det}\left(\lambda S_{00}-S_{01} S_{11}^{-1} S_{10}\right)$ as opposed to the standard eigenvalue problem $0=\operatorname{det}\left(\lambda S_{11}-S_{10} S_{00}^{-1} S_{01}\right)$.

### 3.5. Asymptotic Theory for the Test on the Cointegrating Vectors

We now consider the tests on the cointegrating vectors in the rank deficient case when a constant is present in the model. There is now a wide range of possible limit distributions. Only a few of these will be discussed.

The unrestricted model is $\mathrm{H}_{c}(r)$ where the constant is restricted to the cointegrating space. Thus, in the full rank case the Granger-Johansen representation (22) has a zero linear slope $\tau_{\ell}=0$ and level satisfying $\beta^{\prime} \tau_{c}=-\beta_{c}$.

Consider now the hypothesis of a known cointegrating vector, (21). It is now important whether the hypothesized level for the cointegrating vector, $b_{c}$ is zero or not. If $b_{c} \neq 0$ then a nuisance parameter
depending on $b, b_{c}$ would appear in the limit distributions in the rank deficient case. If $b_{c}=0$ then the limit distributions are simpler. Fortunately, the zero level case is the most natural hypothesis in most applications. The asymptotic theory for the test statistic is described in the following theorems.

Theorem 9. Consider the hypothesis $\mathrm{H}_{c, \beta}(r):(\Pi, \mu)=\alpha b^{* \prime}$ where $b^{*}=\left(b^{\prime}, b_{c}^{\prime}\right)^{\prime}$. Here, $\alpha, b$ have dimension $p \times r$ while $b_{c}^{\prime}$ is an $r$-vector, where $\alpha$ is unknown and $b^{*}$ is known and $b$ has full column rank. Suppose $\mathrm{H}_{z}(0)$ is satisfied so that $\Pi=0, \mu=0$, and $s=0$ and that the $I(1)$ condition is satisfied. Let Be be $p$-dimensional standard Brownian motion on $[0,1]$, where the first $r$ components are denoted $B_{1}$. Define the $(p-s+1)$-dimensional process $F_{u}=\left(B_{u}^{\prime}, 1\right)$ as in (23). Then it holds, for $T \rightarrow \infty$, that

$$
\begin{align*}
& L R\left\{\mathrm{H}_{z, \beta}(r) \mid \mathrm{H}_{z}(p)\right\} \xrightarrow{\mathrm{D}} \operatorname{tr}\left\{\int_{0}^{1} d B_{u} F_{u}^{\prime}\left(\int_{0}^{1} F_{u} F_{u}^{\prime} d u\right)^{-1} \int_{0}^{1} F_{u}\left(d B_{u}\right)^{\prime}\right. \\
&\left.-\int_{0}^{1} d B_{u} B_{1, u}^{\prime}\left(\int_{0}^{1} B_{1, u} B_{1, u}^{\prime} d u\right)^{-1} \int_{0}^{1} B_{1, u}\left(d B_{u}\right)^{\prime}\right\} \tag{24}
\end{align*}
$$

The convergence of the test statistic $\operatorname{LR}\left\{\mathrm{H}_{c, \beta}(r) \mid \mathrm{H}_{c}(p)\right\}$ holds jointly with the convergence for the rank test statistic $\operatorname{LR}\left\{\mathrm{H}_{c}(r) \mid \mathrm{H}_{c}(p)\right\}$, for $s=0$, in Theorem 8. Thus, when $s=0$ a formula of the type (9) implies that the limit distribution of the test statistic for known $\beta$ within the model with rank of at most $r$ satisfies can be found as the difference of the two limiting variables.

Table 7 reports the asymptotic distribution of the test for known cointegrating vector in the model where the rank is at most $r$. When $s=r$, the asymptotic distribution is $\chi^{2}$ with $r(p+1-r)$ degrees of freedom, see Johansen and Juselius (1990, p. 193-194), Johansen et al. (2000, Lemma A.5). When $s=0$ the distribution is simulated according to Theorem 9. It is shifted to the right relative to the case $s=r$.

Table 8 reports the simulated asymptotic distribution of the test for known cointegrating vector in the model where the rank is unrestricted. The distribution is shifted to the right in the rank deficient case. As in the zero level case, the expectations reported in Tables 6 and 7 add up to the expectation reported in Table 8. In the full rank case $s=r$ the statistics in Tables 6 and 7 are independent, as proved below, so the variances are additive.

Theorem 10. Consider the hypothesis $\mathbf{H}_{c, \beta}^{\circ}(r)$. Suppose $\mathrm{H}_{c}^{\circ}(r)=\mathrm{H}_{c}^{\circ}(r) / \mathrm{H}_{c \ell}^{\circ}(r-1)$ is satisfied and that the $I(1)$ condition holds with $s=r$. Then the rank test statistic $L R\left\{\mathrm{H}_{c}^{\circ}(r) \mid \mathrm{H}_{c}^{\circ}(p)\right\}$ and the statistic $L R\left\{\mathrm{H}_{c, \beta}^{\circ}(r) \mid \mathrm{H}_{c}^{\circ}(r)\right\}$ for testing a simple hypothesis on the cointegrating vector are asymptotically independent. The asymptotic distribution of the rank statistic $\operatorname{LR}\left\{\mathrm{H}_{c}^{\circ}(r) \mid \mathrm{H}_{c}^{\circ}(p)\right\}$ is given in Theorem 1, while the statistic for the cointegrating vector $L R\left\{\mathbf{H}_{c, \beta}^{\circ}(r) \mid \mathrm{H}_{c}^{\circ}(r)\right\}$ is asymptotically $\chi^{2}\{r(p+1-r)\}$.

Table 7. Quantiles, mean, and variance of $L R\left\{\mathrm{H}_{c, \beta}(r) \mid \mathrm{H}_{c}(r)\right\}$, where the data generating process satisfies $\mathrm{H}_{c, \beta}^{\circ}(s)$.

| $p$ | $r$ | $s$ | $\mathbf{5 0 \%}$ | $\mathbf{8 0 \%}$ | $\mathbf{8 5 \%}$ | $\mathbf{9 0 \%}$ | $\mathbf{9 5 \%}$ | $\mathbf{9 7 . 5 \%}$ | $\mathbf{9 9 \%}$ | Mean | Var |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 2 | 1 | 1 | 1.39 | 3.22 | 3.79 | 4.61 | 5.99 | 7.38 | 9.21 | 2 | 4 |
|  |  | 0 | 6.34 | 9.84 | 10.78 | 12.02 | 14.05 | 15.96 | 18.41 | 6.87 | 15.09 |
| 3 | 2 | 2 | 3.36 | 5.99 | 6.75 | 7.78 | 9.49 | 11.14 | 13.28 | 4 | 8 |
|  |  | 0 | 12.45 | 17.48 | 18.79 | 20.53 | 23.26 | 25.76 | 28.91 | 13.12 | 30.71 |
| 3 | 1 | 1 | 2.37 | 4.64 | 5.32 | 6.25 | 7.82 | 9.35 | 11.35 | 3 | 6 |
|  |  | 0 | 10.60 | 14.82 | 15.92 | 17.36 | 19.66 | 21.79 | 24.48 | 11.07 | 22.93 |

Table 8. Quantiles, mean, and variance of $L R\left\{\mathrm{H}_{c, \beta}(r) \mid \mathrm{H}_{c}(p)\right\}$, where the data generating process satisfies $\mathrm{H}_{c, \beta}^{\circ}(s)$.

| $p$ | $r$ | $s$ | $\mathbf{5 0 \%}$ | $\mathbf{8 0 \%}$ | $\mathbf{8 5 \%}$ | $\mathbf{9 0 \%}$ | $\mathbf{9 5 \%}$ | $\mathbf{9 7 . 5 \%}$ | $\mathbf{9 9 \%}$ | Mean | Var |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 2 | 1 | 1 | 5.44 | 8.50 | 9.34 | 10.50 | 12.38 | 14.17 | 16.50 | 6.07 | 10.98 |
|  |  | 0 | 9.32 | 13.37 | 14.44 | 15.88 | 18.18 | 20.31 | 22.98 | 9.94 | 19.72 |
| 3 | 2 | 2 | 7.44 | 11.02 | 11.99 | 13.29 | 15.37 | 17.37 | 19.88 | 8.09 | 15.09 |
|  |  | 0 | 15.37 | 20.48 | 21.80 | 23.54 | 26.26 | 28.78 | 31.88 | 15.99 | 32.22 |
| 3 | 1 | 1 | 14.46 | 19.08 | 20.28 | 21.88 | 24.39 | 26.74 | 29.64 | 15.10 | 25.77 |
|  |  | 0 | 20.35 | 25.89 | 27.31 | 29.15 | 32.04 | 34.72 | 38.02 | 20.96 | 38.07 |

## 4. Applications of Results

We discuss how our results apply to the finite sample theory and to identification robust inference. An application to US treasury yields is given.

### 4.1. Finite Sample Theory

The finite sample distribution of cointegration rank tests have been studied in various ways. When there are no nuisance parameters, the asymptotic distributions generally give good approximations. An example is the test for a unit root in a first order autoregression, where the finite sample distribution and the asymptotic distribution are nearly indistinguishable for $T=8$ observations, see Nielsen (1997). A Bartlett correction improves the asymptotic distribution further. Once there are nuisance parameters the situation is different. Under the rank hypothesis the asymptotic distribution differs if there are additional unit roots. This arises either with rank deficiency like here where the distributions tend to be shifted to the left and when there are double roots as in $I(2)$ systems where the distributions are shifted to the right. Nielsen (2004) analyzed this through simulation and suggested to apply local-to-unity approximation that would average between the different asymptotic distributions. A similar idea was implemented analytically for canonical correlation models in Nielsen (1999). In a follow-up paper, Nielsen (2001) analyzed the effects of plugging parameter estimates into such corrections. Johansen (2002) suggested a Bartlett correction for such models. This works quite well when the nuisance parameters are such that they are far from giving additional unit roots. The issue is that the Bartlett correction asymptotes to infinity when there are additional unit roots. More recently, bootstrap methods have been explored by Swensen (2004) and by Cavaliere et al. (2012).

Johansen (2000) derives a Bartlett-type correction for the tests on the cointegrating relations. In Table 2 he considers the finite sample properties of a test comparing the test statistic $L R\left\{\mathrm{H}_{z, \beta}(1) \mid L R\left\{\mathrm{H}_{z}(1)\right\}\right.$ with the asymptotic $\chi^{2}$-approximation. Null rejection frequencies are simulated for dimensions $p=2,5$, a variety of parameter values, and a finite sample size $T$. In all the reported simulations the data generating process has rank of unity. The table shows that null rejection frequency can be very much larger for a nominal $5 \%$ test when the rank is nearly deficient.

Theorem 2 sheds some light on the behaviour of the test as the rank approaches deficiency. The Theorem shows that the test statistic converges for all deficient ranks. Table 2 indicates that the distribution shifts to the right in the rank deficient case. Thus, we should expect that null rejection frequency increases as the rank approaches deficiency, but it should be bounded away from unity.

### 4.2. Identification Robust Inference

Khalaf and Urga (2014) were concerned with tests on cointegation vectors in situations where the cointegration rank is nearly deficient. Their results can be developed a little further using the present results.

The notation in Khalaf and Urga (2014) differs slightly from the present notation. The hypothesis of known cointegration vectors is stated as $\beta_{0}=\left(I_{r}, \mathbf{b}_{0}^{\prime}\right)^{\prime}$ for some known $\mathbf{b}_{0}$, corresponding to the present hypotheses $\mathrm{H}_{z, \beta}(r)$ and $\mathrm{H}_{c \ell, \beta}(r)$. The test statistics are

$$
\begin{align*}
\operatorname{LR}\left(\mathbf{b}_{0}\right) & =\operatorname{LR}\left\{\mathrm{H}_{m, \beta}(r) \mid \mathrm{H}_{m}(p)\right\},  \tag{25}\\
\operatorname{LRC}\left(\mathbf{b}_{0}\right) & =\operatorname{LR}\left\{\mathrm{H}_{m, \beta}(r) \mid \mathrm{H}_{m}(r)\right\} \tag{26}
\end{align*}
$$

for $m=z, c \ell$. Moreover they consider the hypothesis $\mathrm{H}_{m, \Pi}(r)$, say, of a known impact matrix $\Pi$ of rank $r$. This is tested through the statistic

$$
\begin{equation*}
L R_{*}=L R\left\{\mathrm{H}_{m, \Pi}(r) \mid \mathrm{H}_{m}(p)\right\} . \tag{27}
\end{equation*}
$$

When the rank is not deficient the test statistic $\operatorname{LRC}\left(\mathbf{b}_{0}\right)$ is asymptotically $\chi_{r(p-r)}^{2}$, see Johansen (1995, Section 7). The test statistic $L R\left(\mathbf{b}_{0}\right)$ has a Dickey-Fuller type distribution as derived in Theorem 2 for the case without deterministic terms, contradicting the $\chi^{2}$ asymptotics suggested by Khalaf and Urga (2014, Section 4). Table 2 indicates that this distribution is close to, but different from, a $\chi_{p(p-r)}^{2}$-distribution when $p=2,3$ and $p-r=1$. When $p=3$ and $r=1$, the limiting distribution is further from a $\chi_{p(p-r)}^{2}$-distribution. Likewise, the statistic $L R_{*}$ converges to a Dickey-Fuller-type distribution. This can be proved through a modification of the proof of Theorem 2.

Khalaf and Urga's Theorem 1 is concerned with bounding the distribution of the likelihood ratio statistic for the hypothesis $\Pi=a b^{\prime}$, where $a, b$ are known $p \times r$-matrices so that $b$ has rank $r$, against the alternative where $\Pi$ is unrestricted. The idea of their Theorem is to come up with a bound to the critical value when $a, b$ may have deficient rank $s \leq r$. Unfortunately, their theorem evolves around the incorrect $\chi^{2}$ distribution although unit root testing is implicitly involved. We therefore reformulate the result in terms of the limiting distributions derived herein.

We consider the test statistic $L R\left(\mathbf{b}_{0}\right)=L R\left\{\mathrm{H}_{z, \beta}(1) \mid \mathrm{H}_{z}(1)\right\}$ when the rank of $\Pi$ is nearly deficient. Suppose the rank is nearly deficient in the sense that $\Pi \approx T^{-1} M$ for some matrix $M$ along the lines of the theory in Section 2.6. Then, intuitively, the limiting distribution will be a combination of those arising when the true rank is 0 and when it is 1 . The asymptotic theory developed here gives the relevant bounds. In the case of the zero level model the Theorems 1 and 2 imply the following pointwise result.

Theorem 11. Let $\theta$ denote the parameters of the model (1). Consider the parameter space $\Theta_{z}$ where the hypothesis $\mathrm{H}_{z, \beta}(1): \Pi=\alpha b^{\prime}$ holds. Here $\alpha, b$ are both of dimension $p \times 1$. Here $\alpha$ is unknown, while $b$ is known and has full column rank. Suppose the data generating process satisfies the $I(1)$ condition with $s \leq 1$. Let $q_{z, s}$ be the asymptotic $(1-\psi)$ quantile of $\operatorname{LR}\left\{\mathrm{H}_{z, \beta}(1) \mid \mathrm{H}_{z}(1)\right\}$ when the data generating process satisfies $\mathrm{H}_{z, \beta}^{\circ}(s)$ for $s=0,1$. Let $q_{z, *}=\max _{s=0,1} q_{z, s}$. Then it holds for all $\theta \in \Theta_{z}$ that

$$
\begin{equation*}
\lim _{T \rightarrow \infty} \mathrm{P}\left[L R\left\{\mathrm{H}_{z, \beta}(1) \mid \mathrm{H}_{z}(1)\right\} \geq q_{z, *}\right] \leq \psi \tag{28}
\end{equation*}
$$

The simulated values in Table 2 show that for $\psi=5 \%$ then

$$
q_{z, *}=\max \left(q_{z, 0}, q_{z, 1}\right)= \begin{cases}\max (9.05,3.84)=9.05 & \text { for } p=2,  \tag{29}\\ \max (13.82,5.99)=13.82 & \text { for } p=3\end{cases}
$$

The interpretation is as follows. Suppose the hypothesis $\mathrm{H}_{z}(1)$ has not been rejected, but it is unclear whether the rank could be nearly deficient. Then the hypothesis of a known $\beta_{0}$ is rejected if the statistic $\operatorname{LR}\left\{\mathrm{H}_{z, \beta}(1) \mid \mathrm{H}_{z}(1)\right\}$ is larger than $q_{z, *}$.

The bound for $q_{z, *}$ seems very extreme. Khalaf and Urga therefore suggest to use the alternative statistic $\operatorname{LR}\left\{\mathrm{H}_{z, \beta}(1) \mid \mathrm{H}_{z}(p)\right\}$. Theorem 11 could be modified to cover this statistic. The simulations in Table 3 indicate that we would then use bounds

$$
\tilde{q}_{z, *}=\max \left(\tilde{q}_{z, 0}, \tilde{q}_{z, 1}\right)= \begin{cases}\max (9.70,6.22)=9.70 & \text { for } p=2  \tag{30}\\ \max (20.83,15.34)=20.83 & \text { for } p=3\end{cases}
$$

We can establish a similar result for the constant level model using Theorems 8 and 9 . However, it is necessary to exclude the possibility of a linear trends in the rank deficient model as this would give a very complicated result.

Theorem 12. Let $\theta$ denote the parameters of the model (14). Consider the parameter space $\Theta_{c}$ where the hypothesis $\mathrm{H}_{c, \beta}(1):(\Pi, \mu)=\alpha\left(b^{\prime}, b_{c}^{\prime}\right)$ holds. Here $\alpha, b$ are both of dimension $p \times 1$, while $b_{c}$ is a scalar. Further $b, b_{c}$ are known and $b \neq 0$. Suppose the data generating process satisfies the $I(1)$ condition with $s=0$ or $s=1$. Let $q_{c, s}$ be the asymptotic $(1-\psi)$ quantile of $L R\left\{\mathrm{H}_{c, \beta}(1) \mid \mathrm{H}_{c}(1)\right\}$ when the data generating process satisfies $H_{c, \beta}^{\circ}(s)$ for $s=0,1$. Let $q_{c, *}=\max _{s=0,1} q_{c, s}$. Then it holds for all $\theta \in \Theta_{1}$ that

$$
\begin{equation*}
\lim _{T \rightarrow \infty} \mathrm{P}\left[\operatorname{LR}\left\{\mathrm{H}_{c, \beta}(1) \mid \mathrm{H}_{c}(1)\right\} \geq q_{c, *}\right] \leq \psi \tag{31}
\end{equation*}
$$

The simulated values in Table 7 show that for $\psi=5 \%$ then

$$
q_{c, *}=\max \left(q_{z, 0}, q_{z, 1}\right)= \begin{cases}\max (14.05,5.99)=14.05 & \text { for } p=2,  \tag{32}\\ \max (19.66,7.82)=19.66 & \text { for } p=3\end{cases}
$$

If the alternative is taken as $\mathrm{H}_{c}(p)$ instead of $\mathrm{H}_{c}(1)$ the bounds are modified as

$$
\tilde{q}_{c, *}=\max \left(\tilde{q}_{c, 0}, \tilde{q}_{c, 1}\right)= \begin{cases}\max (18.18,12.38)=18.18 & \text { for } p=2  \tag{33}\\ \max (32.04,24.39)=32.04 & \text { for } p=3\end{cases}
$$

The bounds (32), (33) for the constant level model appear further apart than the corresponding bounds (29), (30) for the zero level model. So in the constant level case there is perhaps less reason to use the test against the unrestricted model.

### 4.3. Empirical Illustration

The identification robust inference can be illustrated using a series of monthly US treasury zero-coupon yields over the period 1987:8 to 2000:12. The data are taken from Giese (2008) and runs from the start of Alan Greenspan's chairmanship of the Fed and finishes before the burst of the dotcom bubble. Giese considers 5 maturities ( $1,3,18,48,120$ months), but here we only consider 2 maturities (12, 24 months). The empirical analysis uses OxMetrics, see Doornik and Hendry (2013).

Figure 1 shows the data in levels and differences along with the spread. The spread does not appear to have much of a mean reverting behaviour. It is not crossing the long-run average for periods of up to 4 years. This point towards a random walk behaviour which contradicts the expectations hypothesis in line with Giese's analysis. She finds two common trends among five maturities. The two common trends can be interpreted as short-run and long-run forces driving the yield curve. The cointegrating relations match an extended expectations hypothesis where spreads are not cointegrated but two spreads cointegrate. This is sometimes called butterfly spreads and gives a more flexible match to the yield curve. This is in line with earlier empirical work. Hall et al. (1992), among others, found only one common trend when looking at short-term maturities, while Shea (1992); Zhang (1993) and Carstensen (2003) found more than one common trend when including longer maturities.


Figure 1. Zero coupon yields in (a) levels; (b) differences; and (c) spread.
A vector autoregression of the form (14) with an intercept, $k=4$ lags as well as a dummy variable for 1987:10 was fitted to the data. This has the form

$$
\Delta X_{t}=\Pi X_{t-1}+\mu+\sum_{i=1}^{3} \Gamma_{i} \Delta X_{t-i}+\Phi 1_{(t=1987: 10)}+\varepsilon_{t} \quad \text { for } t=1, \ldots, T
$$

where $X_{t}$ is the bivariate vector of the 12 and 24 month zero-coupon yields and periods $t=1$ and $t=T$ correspond to 1987:8 and 2000:12 giving $T=161$.

Table 9 reports specification test statistics with $p$-values in square brackets. The tests do not provide evidence against the initial model. They are the autocorrelation test of Godfrey (1978) the cumulant based normality test, see Doornik and Hansen (2008), and the ARCH test of Engle (1982). For the validity of applying the autoreregressive and normality tests for non-stationarity autoregressions, see Engler and Nielsen (2009), Kilian and Demiroglu (2000), and Nielsen (2006).

The dummy variable matches the policy intervention after the stock market crash on 19 October 1987. Empirically, the dummy variable can be justified in two ways. First, the plot of yield differences in Figure 1b indicate a sharp drop in yields at that point. Secondly, the robustified least squares algorithm analyzed in Johansen and Nielsen (2016) could be employed for each of the two equations in the model. The algorithm uses a cut-off for outliers in the residuals that is controlled in terms of the gauge, which is the frequency of falsely detected outliers that can be tolerated. The gauge is chosen small in line with recommendations of Hendry and Doornik (2014, Section 7.6), see also Johansen and Nielsen (2016). Thus, we choose a cut-off of 3.02 corresponding to a gauge of $0.25 \%$. When running the autoregressive distributed lag models without outliers, only 1987:10 has an absolute residual exceeding the cut-off. Next, when re-running the model including a dummy for 1987:10, no further residuals exceed the cut-off. This is a fixed point for the algorithm. The detection of outliers may have some impact on specification tests, estimation, and inference. Johansen and Nielsen $(2009,2016)$ analyze the impact on estimation when the data generating process has no outliers. They find that outlier detection only gives a modest efficiency loss compared to standard least squares when the cut-off is as large as chosen here. Berenguer-Rico and Nielsen (2017) find a considerable impact on the normality test employed above. At present, there is no theory for these algorithms for data generating processes with
outliers, albeit some results are available for cointegration analysis with known break date, including the broken trend analysis of Johansen et al. (2000) and the structural change model of Hansen (2003).

Table 9. Specification tests for the unrestricted vector autoregression.

| Test | $b_{12, t}$ | $b_{24, t}$ | Test | System |
| :--- | :---: | :---: | :--- | :---: |
| $\chi_{\text {normality }}^{2}(2)$ | 3.8 | 4.1 | $\chi_{\text {normality }}^{2}(4)$ | 4.3 |
| $F_{\text {ar,1-7 }}(7,144)$ | $[0.15]$ | $[0.13]$ | 1.0 | $F_{\text {ar,1-7 }}(28,272)$ |
| $F_{\text {arch }, 1-7}(7,147)$ | $[0.11]$ | $[0.45]$ | 1.2 |  |
|  | $[0.8$ | 1.0 | $[0.41]$ |  |

Table 10 reports cointegration rank tests. The fifth column shows conventional $p$-values based on Tables 4 and 6 for $s=r$ corresponding to Johansen (1995, Tables 15.2, 15.3). The sixth column shows $p$-values based on Tables 5 and 6 assuming data have been generating by a model satisfying $\mathrm{H}_{c}(0)=\mathrm{H}_{z}(0)$. In both cases the $p$-values are approximated by fitting a Gamma distribution to the reported mean and variance, see Nielsen (1997); Doornik (1998) for details. As expected, the latter $p$-values tend to be higher than the former. Overall this provide overwhelming evidence in favour of a pure random walk model in line with Giese (2008).

Table 10. Cointegration rank tests.

| Hypothesis | $r$ | Likelihood | $L R$ | $p$-Value |  |
| :--- | :--- | :--- | :--- | :--- | :--- |
|  |  |  |  | $s$ | $\mathbf{H}_{c}(\mathbf{0})$ |
| $\mathrm{H}_{c l}(2)=\mathrm{H}_{c}(2)$ | 2 | 134.63 |  |  |  |
| $\mathrm{H}_{C l}(1)$ | 1 | 133.71 | 1.8 | 0.18 | 0.39 |
| $\mathrm{H}_{c}(1)$ | 1 | 133.71 | 1.8 | 0.80 | 0.75 |
| $\mathrm{H}_{c l}(0)$ | 0 | 129.70 | 9.8 | 0.30 | 0.46 |
| $\mathrm{H}_{c}(0)$ | 0 | 129.21 | 10.8 | 0.57 | 0.57 |

If we have a strong belief in the expectation hypothesis we would, perhaps, ignore the rank tests and seek to test the expectations hypothesis directly. If we maintain the model $\mathrm{H}_{c}(1)$, we could have to contemplate that the cointegration vectors could be nearly unidentified. A mild form of the expectation hypothesis is that the spread is zero mean stationary. Thus, we test the restriction $b^{*}=(1,-1,0)$. The likelihood ratio statistic is 4.0. Assuming the data generating process satisfies either $\mathbf{H}_{c}^{\circ}(0)$ or $\mathbf{H}_{c}^{\circ}(1)$, but not by $\mathrm{H}_{c \ell}^{\circ}(0)$, we can apply the Khalaf-Urga (2014)-type bound test established in Theorem 12. The $95 \%$ bound in (32) is 14.05 so the hypothesis cannot be rejected based on this test. This contrasts with the above rank tests which gave strong evidence against the expectations hypothesis. The results reconcile if the bounds test does not have much power in the weakly identified case. Indeed, this seems to be the case when looking at Table 3, $\rho=0.99$-panels in Khalaf and Urga (2014), corresponding to near rank deficiency or weak identification. Thus, assuming the rank is one when in fact the data generating process appears to be nearly rank deficient seems to reduce power for tests on the cointegrating vector. That is, when the alleged cointegrating vector is not cointegrating it would be useful to be able to falsify the economic hypothesis. The above mentioned simulations indicate that this is not the case.

## 5. Conclusions

We have derived asymptotic theory for cointegration rank tests and tests on cointegrating vectors in the rank deficient case. The asymptotic distributions have been simulated and tabulated. The results shed some light on the finite sample theory for cointegration analysis. They can be used to improve the theory on identification robust inference developed by Khalaf and Urga (2014). This was applied to two US treasury yield series.

It appears that large distortions arise when applying standard cointegration inference in the situation where the rank is deficient or nearly deficient. The rank hypothesis gives an inequality for the rank, that is rank $\Pi \leq r$. This includes cases where the rank is $r$ and where it is less than $r$. Thus, the parameter space for the model where rank $\Pi \leq r$ therefore has a lower dimensional subset where the rank is deficient. Inferential procedures for rank determination are consistent but do leave a positive probability of deciding for a deficient rank in finite samples. In practice, it is therefore possible to end up in a situation of rank deficiency or near deficiency. When proceeding to testing restrictions on the cointegrating vectors, the model is therefore mis-specified or nearly mis-specified.

The asymptotic analysis of the test distributions gives the following results. When testing for cointegration rank, the distribution shifts to the left when the rank is deficient. When testing for restrictions on the cointegrating vector, the distribution shifts to the right when the rank is deficient. When the rank is nearly deficient the distribution will tend to shift in similar directions. As a consequence, a test for cointegration restrictions using conventional critical has a size control problem previously observed by Johansen (2000). One can instead apply identification robust tests as suggested by Khalaf and Urga (2014), but our impression is that while these tests are better behaved in terms of size, they have modest power to reject incorrect restrictions.

Our recommendation is to test for rank before testing restrictions on cointegrating vectors in line with Johansen's framework. If the conclusion from the rank determination is ambiguous it is best to proceed with caution and possibly explore different choices for rank. This is a common theme in the applied work of Juselius.

Author Contributions: The authors made equal contributions.
Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.

## Appendix A. Proofs

Processes are considered on the space of right continuous processes with left limits, $D[0,1]$. A discrete time process $X_{t}$ for $t=1, \ldots, T$ is embedded in $D[0,1]$ through $X_{\text {integer }(T u)}$ for $0 \leq u \leq 1$. For processes $Y_{t}, Z_{t}$ for $t=1, \ldots, T$ the residuals from regressing $Y_{t}$ on $Z_{t}$ are denoted $\left(Y_{t} \mid Z_{t}\right)=$ $Y_{t}-\sum_{s=1}^{T} Y_{s} Z_{s}^{\prime}\left(\sum_{s=1}^{T} Z_{s} Z_{s}^{\prime}\right)^{-1} Z_{t}$.

Proof of Theorem 1. This follows the outline of the proof in Johansen (1995, §10, 11). Let $\Pi=\alpha_{0} \beta_{0}^{\prime}$ for $p \times s$-matrices $\alpha_{0}, \beta_{0}$ with full column rank. Let $\Gamma=I_{p}-\sum_{i=1}^{k-1} \Gamma_{i}$. Under the $\mathrm{I}(1)$ condition the Granger-Johansen representation (6) holds with rank $s$ and Johansen's Lemma 10.1 stands with $r$ replaced by $s$. His Lemmas 10.2, 10.3 hold with $B_{T}=\beta_{0 \perp}\left(\beta_{0 \perp}^{\prime} \beta_{0 \perp}\right)^{-1}$ so that, on $D[0,1]$,

$$
\begin{equation*}
T^{-1 / 2} B_{T}^{\prime} X_{\mathrm{integer}(T u)}=B_{T}^{\prime} C T^{-1 / 2} \sum_{t=1}^{\operatorname{integer}(T u)} \varepsilon_{t}+\mathrm{op}(1) \tag{A1}
\end{equation*}
$$

For later use we will note that the Brownian motion $B$ can be chosen as follows. For any orthogonal square matrix $\tilde{M}$ so $\tilde{M}^{\prime} \tilde{M}=I_{p-s}$ choose the $(p-s)$-dimensional standard Brownian motion $B$ so that

$$
\begin{equation*}
T^{-1 / 2} \tilde{M}^{\prime}\left(\alpha_{0 \perp}^{\prime} \Omega \alpha_{0 \perp}\right)^{-1 / 2} \alpha_{0 \perp}^{\prime} \Gamma \beta_{0 \perp}\left(\beta_{0 \perp}^{\prime} \beta_{0 \perp}\right)^{-1} \beta_{0 \perp}^{\prime} X_{[T u]} \xrightarrow{\mathrm{D}} B_{u} \tag{A2}
\end{equation*}
$$

on $D[0,1]$.

Proof of Theorem 2. Introduce the notation $\widehat{\Omega}_{U}=S_{00}-S_{01} S_{11}^{-1} S_{10}$ for the unrestricted variance estimator and $\widehat{\Omega}_{R}=S_{00}-S_{01} b\left(b^{\prime} S_{11} b\right)^{-1} b^{\prime} S_{10}$ for the restricted variance estimator. Then the likelihood ratio test statistic satisfies

$$
L R\left\{\mathrm{H}_{z, \beta}(r) \mid \mathrm{H}_{z}(p)\right\}=-T \log \frac{\operatorname{det}\left(\widehat{\Omega}_{U}\right)}{\operatorname{det}\left(\widehat{\Omega}_{R}\right)}=T \log \operatorname{det}\left\{I_{p}+\widehat{\Omega}_{U}^{-1}\left(\widehat{\Omega}_{R}-\widehat{\Omega}_{U}\right)\right\}
$$

If it is shown that $\widehat{\Omega}_{U}$ is consistent and $T\left(\widehat{\Omega}_{R}-\widehat{\Omega}_{U}\right)$ converges in distribution then

$$
\begin{equation*}
L R\left\{\mathrm{H}_{z, \beta}(r) \mid \mathrm{H}_{z}(p)\right\}=\operatorname{tr}\left\{\Omega^{-1} T\left(\widehat{\Omega}_{R}-\widehat{\Omega}_{U}\right)\right\}+\mathrm{op}(1) \tag{A3}
\end{equation*}
$$

following Johansen (1995, p. 224). The consistency of the unrestricted variance estimator $\widehat{\Omega}_{U}$ follows from Johansen (1995, Lemma 10.3) used with $r=s=0$ and $B_{T}=I_{p}$.

Consider $T\left(\widehat{\Omega}_{R}-\widehat{\Omega}_{U}\right)$. Note first that the data generating process has cointegration rank $s=0$. Thus $\alpha_{0}, \beta_{0}$ are empty matrices so that their complements can be chosen as the identity matrix. The I(1) condition then implies that $\Gamma=I_{p}-\sum_{i=1}^{k-1} \Gamma_{i}$ is invertible. The asymptotic convergence in (A2) then reduces to

$$
\begin{equation*}
T^{-1 / 2} \tilde{M}^{\prime} \Omega^{-1 / 2} \Gamma X_{\text {integer }(T u)}=T^{-1 / 2} \tilde{M}^{\prime} \Omega^{-1 / 2} \sum_{t=1}^{\text {integer }(T u)} \varepsilon_{t}+\mathrm{op}(1) \xrightarrow{\mathrm{D}} B_{u} \tag{A4}
\end{equation*}
$$

where $B$ is a standard Brownian motion of dimension $p$ and for any orthonormal $\tilde{M}$ so that $\tilde{M}^{\prime} \tilde{M}=I_{p}$. In particular, we will choose $\tilde{M}$ so

$$
\tilde{M}=\left[\begin{array}{c}
\left\{b^{\prime} \Gamma^{-1} \Omega\left(\Gamma^{\prime}\right)^{-1} b\right\}^{-1 / 2} b^{\prime} \Gamma^{-1} \Omega^{1 / 2}  \tag{A5}\\
\left(b_{\perp}^{\prime} \Gamma^{\prime} \Omega^{-1} \Gamma b_{\perp}\right)^{-1 / 2} b_{\perp}^{\prime} \Gamma^{\prime} \Omega^{-1 / 2}
\end{array}\right]
$$

Let $B_{1, u}, B_{2, u}$ be the first $r$ and the last $p-r$ coordinates of $B_{u}$, respectively. Then we get

$$
\left\{b^{\prime} \Gamma^{-1} \Omega\left(\Gamma^{\prime}\right)^{-1} b\right\}^{-1 / 2} b^{\prime} X_{\text {integer }(T u)} \xrightarrow{\mathrm{D}} B_{1, u} .
$$

The variance estimators are $\widehat{\Omega}_{R}=S_{\varepsilon \varepsilon}-S_{\varepsilon 1} b\left(b^{\prime} S_{11} b\right)^{-1} b^{\prime} S_{1 \varepsilon}$ and $\widehat{\Omega}_{U}=S_{\varepsilon \varepsilon}-S_{\varepsilon 1} S_{11}^{-1} S_{1 \varepsilon}$. In particular, the difference of the variance estimators is

$$
\begin{equation*}
T\left(\widehat{\Omega}_{R}-\widehat{\Omega}_{U}\right)=T\left\{S_{\varepsilon 1} M\left(M^{\prime} S_{11} M\right)^{-1} M^{\prime} S_{1 \varepsilon}-S_{\varepsilon 1} b m\left(m^{\prime} b^{\prime} S_{11} b m\right)^{-1} m b^{\prime} S_{1 \varepsilon}\right\} \tag{A6}
\end{equation*}
$$

for any invertible matrices $M, m$ and in particular for $M^{\prime}=\tilde{M}^{\prime} \Omega^{-1 / 2} \Gamma$ and $m=\left\{b^{\prime} \Gamma^{-1} \Omega\left(\Gamma^{\prime}\right)^{-1} b\right\}^{-1 / 2}$. In light of the identity $\tilde{M}^{\prime} \tilde{M}=I_{p}$, the random walk convergence in (A4), the rules for the trace and the notation $v=m b$ write

$$
\begin{aligned}
\operatorname{tr}\left\{\Omega^{-1} T\left(\widehat{\Omega}_{R}-\widehat{\Omega}_{U}\right)\right\} & =\operatorname{tr}\left\{\tilde{M}^{\prime} \Omega^{-1 / 2} T\left(\widehat{\Omega}_{R}-\widehat{\Omega}_{U}\right) \Omega^{-1 / 2} \tilde{M}\right\} \\
& =\operatorname{tr}\left[\tilde{M}^{\prime} \Omega^{-1 / 2} T\left\{S_{\varepsilon 1} M\left(M^{\prime} S_{11} M\right)^{-1} M^{\prime} S_{1 \varepsilon}-S_{\varepsilon 1} v\left(v^{\prime} S_{11} v\right)^{-1} v^{\prime} S_{1 \varepsilon}\right\} \Omega^{-1 / 2} \tilde{M}\right]
\end{aligned}
$$

Then the product moment convergence results in Johansen (1995, Lemma 10.3) imply

$$
\begin{aligned}
& \operatorname{tr}\left\{\Omega^{-1} T\left(\widehat{\Omega}_{R}-\widehat{\Omega}_{U}\right)\right\} \xrightarrow{\mathrm{D}} \operatorname{tr}\left\{\int_{0}^{1} d B_{u} B_{u}^{\prime}\left(\int_{0}^{1} B_{u} B_{u}^{\prime} d u\right)^{-1} \int_{0}^{1} B_{u}\left(d B_{u}\right)^{\prime}\right. \\
&\left.-\int_{0}^{1} d B_{u} B_{1, u}^{\prime}\left(\int_{0}^{1} B_{1, u} B_{1, u}^{\prime} d u\right)^{-1} \int_{0}^{1} B_{1, u}\left(d B_{u}\right)^{\prime}\right\}
\end{aligned}
$$

This is also the limit of the likelihood ratio test statistic due to (A3). The convergence holds jointly with the convergence of the likelihood ratio test statistic for rank in Theorem 1 since the orthogonal matrix $\tilde{M}$ in (A2) can be chosen freely.

Proof of Theorem 3. We need a number of results from Johansen (1995). Let $B, V$ be independent standard Brownian motions. His Theorem 11.1 shows

$$
\begin{equation*}
L R\left\{\mathrm{H}_{z}(r) \mid \mathrm{H}_{z}(p)\right\} \xrightarrow{\mathrm{D}} \operatorname{tr}\left\{\int_{0}^{1} d B_{u} B_{u}^{\prime}\left(\int_{0}^{1} B_{u} B_{u}^{\prime} d u\right)^{-1} \int_{0}^{1} B_{u} d B_{u}^{\prime}\right\} \tag{A7}
\end{equation*}
$$

while his Lemma 13.8 shows

$$
\begin{equation*}
L R\left\{\mathrm{H}_{z, \beta}(r) \mid \mathrm{H}_{z}(r)\right\} \xrightarrow{\mathrm{D}} \operatorname{tr}\left\{\int_{0}^{1} d V_{u} B_{u}^{\prime}\left(\int_{0}^{1} B_{u} B_{u}^{\prime} d u\right)^{-1} \int_{0}^{1} B_{u} d V_{u}^{\prime}\right\} . \tag{A8}
\end{equation*}
$$

Johansen does not explicitly argue that the convergence results hold jointly. This can be done by going into the proofs of the results, find the asymptotic expansions of the test statistic, and express them in terms of random walks that converge to the processes $B, V$ when normalized by $T^{1 / 2}$. The asymptotic distribution in (A8) is mixed Gaussian since $B, V$ are independent. Thus, by conditioning on $B$ we see that $L R\left\{\mathrm{H}_{z, \beta}(r) \mid \mathrm{H}_{z}(r)\right\}$ is asymptotically $\chi^{2}$ and hence independent of $B$. In turn the two test statistics are asymptotically independent.

Proof of Theorem 5. We follow Stockmarr and Jacobsen (1994) or Johansen (1995, Theorem 14.1, Lemma 14.3) and find that $T^{-1 / 2} X_{\text {integer }(T u)}$ converges to $J_{u}$ as a process on $D[0,1]$ while $\left(S_{00}, S_{1 \varepsilon}, S_{11} / T\right)$ converges in distribution to $\left(I_{2}, \int_{0}^{1} J_{u} d B_{u}^{\prime}, \int_{0}^{1} J_{u} J_{u}^{\prime} d u\right)$.

Now, proceed as in the proof of Theorem 2. It has to be argued that $\widehat{\Omega}_{U}$ converges in probability to $I_{2}$ and that $T\left(\widehat{\Omega}_{R}-\widehat{\Omega}_{U}\right)$ has the limit distribution postulated in the Theorem. The convergence of the $\widehat{\Omega}_{U}$ follows from the listed properties of the product moment matrices. For $T\left(\widehat{\Omega}_{R}-\widehat{\Omega}_{U}\right)$ we have as in Equation (A6) that

$$
T\left(\widehat{\Omega}_{R}-\widehat{\Omega}_{U}\right)=T\left\{S_{\varepsilon 1}\left(S_{11}\right)^{-1} S_{1 \varepsilon}-S_{\varepsilon 1} b\left(b^{\prime} S_{11} b\right)^{-1} b^{\prime} S_{1 \varepsilon}\right\}
$$

Again, we can apply the listed properties of the product moment matrices.
Proof of Theorem 6. Similar to the proof of Theorem 1, the relevant Granger-Johansen representation is (22) with rank $s$. Use Johansen's Lemmas 10.2, 10.3 with $B_{T}=\left\{\gamma\left(\gamma^{\prime} \gamma\right)^{-1}, T^{-1 / 2} \tau_{\ell}\left(\tau_{\ell}^{\prime} \tau_{\ell}\right)^{-1}\right\}$, where $\tau_{\ell}=C \mu$, while $\gamma \in \operatorname{span}\left(\beta_{0 \perp}\right)$ so that $\gamma^{\prime} \tau_{\ell}=0$ and the expansion (A1) is replaced by

$$
T^{-1 / 2} B_{T}^{\prime} X_{\text {integer }(T u)}=\left\{\begin{array}{c}
\left(\gamma^{\prime} \gamma\right)^{-1} \gamma^{\prime} C T^{-1 / 2} \sum_{t=1}^{\text {integer }(T u)} \varepsilon_{t}  \tag{A9}\\
u
\end{array}\right\}+\mathrm{op}_{\mathrm{p}}(1)
$$

on $D[0,1]$. Thus, $\Delta X_{t}$ has a non-zero level, but this is eliminated by regression on the intercept.
Proof of Theorem 7. Similar to the proof of Theorem 1. Use the Granger-Johansen representation (22) with rank $s$ and $\tau_{\ell}=C \mu=0$, and Johansen's Lemmas $10.2,10.3$ with $B_{T}=\beta_{0 \perp}\left(\beta_{0 \perp}^{\prime} \beta_{0 \perp}\right)^{-1}$ so that $T^{-1 / 2} B_{T}^{\prime} X_{\text {integer( } T u)}$ has expansion (A1).
Proof of Theorem 8. Similar to the proof of Theorem 1. Use the Granger-Johansen representation (22) with rank $s$, and $\tau_{\ell}$. Use Johansen's Lemmas $10.2,10.3$ with $X_{t}, B_{T}$ and the expansion (A1) replaced by, respectively, $X_{t}^{*}=\left(X_{t}^{\prime}, 1\right)^{\prime}$, the block diagonal matrix $B_{T}^{*}=\operatorname{diag}\left(B_{T}, T^{1 / 2}\right)$ where $B_{T}=\beta_{0 \perp}\left(\beta_{0 \perp}^{\prime} \beta_{0 \perp}\right)^{-1}$, and

$$
\begin{equation*}
T^{-1 / 2} B_{T}^{* \prime} X_{\mathrm{integer}(T u)}^{*}=\binom{B_{T}^{\prime} C T^{-1 / 2} \sum_{t=1}^{\mathrm{integer}(T u)} \varepsilon_{t}}{1}+\mathrm{op}_{\mathrm{P}}(1) \tag{A10}
\end{equation*}
$$

on $D[0,1]$.

Proof of Theorem 9. The proof of Theorem 2 is modified noting that $R_{1, t}$ is the $(p+1)$-vector $\left(X_{t-1}, 1\right)^{\prime}$ corrected for lagged differences instead of $X_{t-1}$ corrected for lagged differences. Choose $\tilde{M}$ as in (A5). Replace (A4) by

$$
\left(\begin{array}{cc}
T^{-1 / 2} \tilde{M}^{\prime} \Omega^{-1 / 2} \Gamma & 0  \tag{A11}\\
0 & 1
\end{array}\right)\binom{X_{\text {integer }(T u)}}{1} \xrightarrow{\mathrm{D}} F_{u} .
$$

The difference of variance estimators in (A6) is now

$$
\begin{equation*}
T\left(\widehat{\Omega}_{R}-\widehat{\Omega}_{U}\right)=T\left\{S_{\varepsilon 1} M\left(M^{\prime} S_{11} M\right)^{-1} M^{\prime} S_{1 \varepsilon}-S_{\varepsilon 1} b^{*}\left(b^{* \prime} S_{11} b^{*}\right)^{-1} b^{* \prime} S_{1 \varepsilon}\right\} \tag{A12}
\end{equation*}
$$

where the invertible $(p+1)$-dimensional matrix $M$ now is chosen as

$$
M=\left\{\begin{array}{ccc}
b^{\prime} \Gamma^{-1} \Omega\left(\Gamma^{\prime}\right)^{-1} b & 0 & 0  \tag{A13}\\
0 & b_{\perp}^{\prime} \Gamma^{\prime} \Omega^{-1} \Gamma b_{\perp} & 0 \\
0 & 0 & 1
\end{array}\right\}^{-1 / 2}\left(\begin{array}{cc}
b^{\prime} & b_{c}^{\prime} \\
b_{\perp}^{\prime} \Gamma^{\prime} \Omega^{-1} \Gamma & 0 \\
0 & 1
\end{array}\right)
$$

Viewed as a $(3 \times 2)$-block matrix, the two upper left equals the previous $M$. Since the random walk dominates a constant it holds that

$$
\left(\begin{array}{cc}
T^{-1 / 2} I_{p} & 0  \tag{A14}\\
0 & 1
\end{array}\right) M\binom{X_{\text {integer }(T u)}}{1} \xrightarrow{\mathrm{D}} F_{u} .
$$

Moreover, the first $r$ coordinates of $M R_{1, t}$ are proportional to $b^{* \prime} R_{1, t}$. Thus the argument can be completed as in the proof of Theorem 2.

Proof of Theorem 10. The proof of Theorem 3 has to be modified to allow for a constant term in the cointegrating vector. The arguments leading to asymptotic results for the test statistics are sketched in Johansen and Juselius (1990) and, with more details, in Johansen et al. (2000, Theorem 3.1, Lemma A.5).

Proof of Theorem 11. Write

$$
\begin{equation*}
L R\left\{\mathrm{H}_{z, \beta}(1) \mid \mathrm{H}_{z}(1)\right\}=\operatorname{LR}\left\{\mathrm{H}_{z}(1) \mid \mathrm{H}_{z}(p)\right\}-L R\left\{\mathrm{H}_{z, \beta}(1) \mid \mathrm{H}_{z}(p)\right\} . \tag{A15}
\end{equation*}
$$

When $s=0$ Theorems 1 and 2 give expansions for the right hand expressions of (A15) and in turn for the desired test statistic on the left hand of (A15). This implies an asymptotic distribution with asymptotic $(1-\psi)$ quantile $q_{z, 0}$, say. When $s=1$ Theorem 3 in a similar way gives an asymptotic $(1-\psi)$ quantile $q_{z, 1}$. Thus, with $q_{z, *}=\max _{s=0,1} q_{z, s}$ we get $\lim _{T \rightarrow \infty} \mathrm{P}\left[L R\left\{\mathrm{H}_{z, \beta}(1) \mid \mathrm{H}_{z}(1)\right\} \geq q_{z, *}\right] \leq \psi$, both with $s=0$ and when $s=1$.

Proof of Theorem 12. Similar to the proof of Theorem 11, applying Theorems 8-10 instead Theorems 1-3.

## References

Abadir, Karim M. 1995. The limiting distribution of the t ratio under a unit root. Econometric Theory 11: 775-93. [CrossRef]
Andrews, Donald W. K., and Xu Cheng. 2012. Estimation and inference with weak, semi-strong, and strong identification. Econometrica 80: 2153-11. [CrossRef]
Berenguer-Rico, Vanessa, and Bent Nielsen. 2017. Marked and Weighted Empirical Processes of Residuals With Applications to Robust Regressions. Discussion Paper 841. Oxford: Department of Economics, University of Oxford.
Bernstein, David. 2014. Asymptotic Theory for Unidentified Cointegration Estimators. M.Phil. thesis, University of Oxford, Oxford, UK.

Carstensen, Kai. 2003. Nonstationary term premia and cointegration of the term structure. Economics Letters 80: 409-13. [CrossRef]
Cavaliere, Giuseppe, Anders Rahbek, and A. M. Robert Taylor. 2012. Bootstrap determination of the co-integration rank in vector autoregressive models. Econometrica 80: 1721-40.
Doornik, Jurgen A. 1998. Approximations to the asymptotic distribution of cointegration tests. Journal of Economic Surveys 12: 573-93. [CrossRef]
Doornik, Jurgen A. 2007. Object-Oriented Matrix Programming Using Ox, 3rd ed. London: Timberlake.
Doornik, Jurgen A., and Henrik Hansen. 2008. An omnibus test for univariate and multivariate normality. Oxford Bulletin of Economics and Statistics 70: 927-39. [CrossRef]
Doornik, Jurgen A. and David F. Hendry. 2013. PcGive 14. London: Timberlake, vol. 1.
Dufour, Jean-Marie. 1997. Some impossibility theorems in econometrics with applications to structural and dynamic methods. Econometrica 65: 1365-87. [CrossRef]
Engle, Robert F. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50: 987-1108. [CrossRef]
Engler, Eric, and Bent Nielsen. 2009. The empirical process of autoregressive residuals. Econometrics Journal 12: 367-81. [CrossRef]
Fachin, Stefano. 2000. Bootstrap and asymptotic tests of long-run relationships in cointegrated systems. Oxford Bulletin of Economics and Statistics 62: 543-51. [CrossRef]
Giese, Julia. 2008. Level, slope, curvature: Characterising the yield curve in a cointegrated VAR model. Economics 2: 28.
Godfrey, L. G. 1978. Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. Econometrica 46: 1293-301. [CrossRef]
Gredenhoff, Mikael, and Tor Jacobson. 2001. Bootstrap testing linear restrictions on cointegrating vectors. Journal of Business \& Economic Statistics 19: 63-72.
Hall, Anthony D., Heather M. Anderson, and Clive W. J. Granger. 1992. A cointegration analysis of treasury bill yields. Review of Economics and Statistics 74: 116-26. [CrossRef]
Hansen, Peter Reinhard. 2003. Structural changes in the cointegrated vector autoregressive model. Journal of Econometrics 114: 261-95. [CrossRef]
Hendry, David F., and Jurgen A. Doornik. 2014. Empirical Model Discovery and Theory Evaluation: Automatic Selection Methods in Econometrics. London: MIT Press.
Johansen, Søren. 1988. Statistical analysis of cointegration vectors. Journal of Economic Dynamics and Control 12: 231-54. [CrossRef]
Johansen, Søren. 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica 59: 1551-580. [CrossRef]
Johansen, Søren. 1992. Determination of cointegration rank in the presence of a linear trend. Oxford Bulletin of Economics and Statistics 54: 383-97. [CrossRef]
Johansen, Søren. 1995. Likelihood Based Inference on Cointegration in the Vector Autoregressive Model. Oxford: Oxford University Press.
Johansen, Søren. 2000. A Bartlett correction factor for tests on the cointegrating relations. Econometric Theory 16: 740-77. [CrossRef]
Johansen, Søren. 2002. A small sample correction of the test for cointegrating rank in the vector autoregressive model. Econometrica 70: 1929-61. [CrossRef]
Johansen, Søren, and Katarina Juselius. 1990. Maximum likelihood estimation and inference on cointegration-With applications to the demand for money. Oxford Bulletin of Economics and Statistics 52: 169-210. [CrossRef]
Johansen, Søren, Rocco Mosconi, and Bent Nielsen. 2000. Cointegration analysis in the presence of structural breaks in the deterministic trend. Econometrics Journal 3: 216-49. [CrossRef]
Johansen, Søren, and Bent Nielsen. 2009. Saturation by indicators in regression models. In The Methodology and Practice of Econometrics: Festschrift in Honour of David F. Hendry. Edited by Jennifer L. Castle and Neil Shephard. Oxford: Oxford University Press, pp. 1-36.
Johansen, Søren, and Bent Nielsen. 2016. Asymptotic theory of outlier detection algorithms for linear time series regression models (with discussion). Scandinavian Journal of Statistics 43: 321-81. [CrossRef]
Juselius, Katarina. 2006. The Cointegrated VAR Model. Oxford: Oxford University Press.
Khalaf, Lynda, and Giovanni Urga. 2014. Identification robust inference in cointegrating regressions. Journal of Econometrics 182: 385-96. [CrossRef]

Kilian, Lutz, and Ufuk Demiroglu. 2000. Residual-based tests for normality in autoregressions: Asymptotic theory and simulation evidence. Journal of Business \& Economic Statistics 18: 40-50.
Mavroeidis, Sophocles, Mikkel Plagborg-Møller, and James H. Stock. 2014. Empirical evidence on inflation expectations in the new Keynesian Phillips curve. Journal of Economic Litterature 52: 124-88. [CrossRef]
Nielsen, Bent. 1997. Bartlett correction of the unit root test in autoregressive models. Biometrika 84: 500-504. [CrossRef]
Nielsen, Bent. 1999. The likelihood ratio test for rank in bivariate canonical correlation analysis. Biometrika 86: 279-88. [CrossRef]
Nielsen, Bent. 2001. Conditional test for rank in bivariate canonical correlation analysis. Biometrika 88: 874-80. [CrossRef]
Nielsen, Bent. 2004. On the distribution of likelihood ratio test statistics for cointegration rank. Econometric Reviews 23: 1-23. [CrossRef]
Nielsen, Bent. 2006. Order determination in general vector autoregressions. In Time Series And Related Topics: In Memory of Ching-Zong Wei. Edited by Hwai-Chung Ho, Ching-Kang Ing and Tze Leung Lai. Lecture Notes-Monograph Series. Beachwood: Institute of Mathematical Statistics, vol. 52, pp. 93-112.
Nielsen, Bent, and Anders Rahbek. 2000. Similarity issues in cointegration models. Oxford Bulletin of Economics and Statistics 62: 5-22. [CrossRef]
Paruolo, Paolo. 2001. The power of lambda max. Oxford Bulletin of Economics and Statistics 63: 395-403. [CrossRef] Phillips, Peter C. B., and Victor Solo. 1992. Asymptotics for linear processes. Annals of Statistics 20: 971-1001. [CrossRef]
Shea, Gary S. 1992. Benchmarking the expectations hypothesis of the interest-rate term structure: An analysis of cointegration vectors. Journal of Business \& Economic Statistics 10: 347-366.
Stockmarr, Anders, and Martin Jacobsen. 1994. Gaussian diffusion and autoregressive processes: Weak convergence and statistical inference. Scandinavian Journal of Statistics 21: 403-19.
Swensen, Anders Rygh 2004. Bootstrap algorithms for testing and determining the cointegration rank in VAR models. Econometrica 74: 1699-714. Corrigendum in volume 77: 1703-704. [CrossRef]
Zhang, Hua. 1993. Treasury yield curves and cointegration. Applied Economics 25: 361-67. [CrossRef]
© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http:/ / creativecommons.org/licenses/by/4.0/).

## Article

# Partial Cointegrated Vector Autoregressive Models with Structural Breaks in Deterministic Terms 

Takamitsu Kurita ${ }^{1, \dagger}$ and Bent Nielsen ${ }^{2, *, \dagger}$<br>1 Faculty of Economics, Fukuoka University, 8-19-1, Nanakuma, Jonan-ku, Fukuoka 814-0180, Japan; tkurita@fukuoka-u.ac.jp<br>2 Department of Economics and Program for Economic Modelling, University of Oxford \& Nuffield College, Oxford OX1 1NF, UK<br>* Correspondence: bent.nielsen@nuffield.ox.ac.uk<br>$\dagger$ These authors contributed equally to this work.

Received: 10 December 2018; Accepted: 30 September 2019; Published: 6 October 2019


#### Abstract

This paper proposes a class of partial cointegrated models allowing for structural breaks in the deterministic terms. Moving-average representations of the models are given. It is then shown that, under the assumption of martingale difference innovations, the limit distributions of partial quasi-likelihood ratio tests for cointegrating rank have a close connection to those for standard full models. This connection facilitates a response surface analysis that is required to extract critical information about moments from large-scale simulation studies. An empirical illustration of the proposed methodology is also provided.


Keywords: partial cointegrated vector autoregressive models; structural breaks; deterministic terms; weak exogeneity; cointegrating rank; response surface

JEL Classification: C12; C32; C50

## 1. Introduction

Partial cointegration models with structural shifts in level or linear trends are quite common in practice; however, no formal analysis is available for these models. The likelihood analysis of the partial models with such breaks is based on reduced rank regression, just like standard full cointegrated vector autoregressive models introduced by Johansen $(1988,1995)$. The main difference lies in the fact that likelihood-based tests for cointegrating rank in the partial models involve a set of new asymptotic distributions which reflect the combination of weakly exogenous regressors and broken deterministic terms. We generalise the standard assumption of normal innovations (Johansen 1995) to a flexible class of heterogeneous martingale difference innovations. We then derive the asymptotic distributions of the test statistics in question and provide a simulated responsed surface of the asymptotic distribution.

The presented models combine two widely used extensions of Johansen's original model. The first extension was a partial cointegrated system investigated by Harbo et al. (1998), referred to as HJNR henceforth, see also Pesaran et al. (2000). This partial system is a conditional vector autoregressive model for a vector of variables, $Y_{t}$, given another vector of variables, $Z_{t}$, as well as lags of both variables. They also presented simulated tables for asymptotic rank test distributions based on the partial system. Boswijk (1995) and Ericsson and MacKinnon (2002) explored the use of conditional autoregressive models. Recently, Cavaliere et al. (2018) considered information criteria based on the HNJR test statistics. The second extension was a full cointegrated system with structural breaks in a constant level or linear trend, a model explored by Johansen et al. (2000), referred to as JMN hereafter. This full model is a multivariate extension of model C of Perron (1989), where both level and linear trend slope change at the time of the break, as opposed to his models A and B, in which only one of the
two is changing. Deterministic breaks in cointegrated systems have also been explored by Inoue (1999) and Hendry and Massmann (2007).

Each of the two extensions above has proved to be useful in empirical applications; furthermore, subsequent practical work has shown that we frequently require both of the two extensions simultaneously. As an example, Bårdsen et al. (2005) built a large scale model of the Norwegian economy by combining a number of smaller partial cointegration models. Each of these sub-systems is regarded as a partial model subject to structural shifts, and these types of models are useful in a practical sense for empirical macroeconomic research. As it stands, however, the exact asymptotic properties of likelihood-based test statistics derived from the partial models with structural breaks are unknown, so that a formal econometric study based on these models is unfeasible. This paper, therefore, conducts both analytical and simulation-based investigations into the unknown asymptotic properties so that researchers can perform a formal analysis using the partial models with structural breaks. Another example of these partial models is a trade model for the UK by Schreiber (2015), which we are going to use as an empirical illustration later in this paper.

This paper shows that the asymptotic distributions of the proposed likelihood-based test statistics are dependent on information about the dimension of the variables $Y_{t}$ and $Z_{t}$, cointegrating rank, the number of breaks and their locations, but the distributions themselves are free of any unknown parameters. Hence, the limit distributions can be simulated given the above information, as in a manner similar to Johansen (1995, §15), HJNR or MacKinnon et al. (1999). The Granger-Johansen representation for the full model in JMN is also reexamined as a basis for the required asymptotic study, and this reexamination can be viewed as a useful clarification of roles of a set of starting values in the workings of the system. It should be noted that a condition for weak exogeneity reviewed in Section 2 is assumed to be satisfied when exploring the properties of the test statistics; the violation of this condition can give rise to a class of limit results that are unfavourable in applications, as discussed by Johansen (1992a). This assumption is testable by following an ex-post testing procedure suggested by Johansen (1992a) and others. We demonstrate this procedure in the empirical illustration in Section 5.

In deriving the asymptotic distributions of the test statistics, the assumption of normal innovations in Johansen (1995), HJNR and JMN, is relaxed to the assumption of martingale difference innovations, with a view to widening the scope of applications of the proposed models. This means we have to be careful in developing asymptotic arguments required for the quasi-likelihood ratio test statistics. We use martingale limit results of Anderson and Kunitomo (1992) and Brown (1971) for approximately stationary components and for non-stationary components, respectively.

Furthermore, it is shown that the derived asymptotic distributions can be approximated by gamma distributions, a class of common statistical distributions identifiable only by the first two moments; the validity of this gamma-distribution approximation method in various other existing models was documented by Nielsen (1997), Doornik (1998) and JMN. The study utilises the fact that mean and variance of the limit distributions for the proposed partial models are expressible in terms of the mean and variance for full models and certain covariance terms. As a result, it is feasible to apply the gamma approximation method to simulation results based on the fullmodels, in order to obtain precise limit quantiles of the test statistics for the proposed partial models. Hence, we are justified in conducting comprehensive simulations in the full-model framework, the results of which are applied in a response surface analysis combined with the gamma approximation method. The outcomes of the response surface analysis are tabulated in two tables, the accuracy of which is verified by moving back to the partial-model framework. The tables allow researchers to conduct formal applied studies with the proposed partial models. A brief empirical study is also provided.

Overall, this paper adds to the literature on time series econometrics and applied macroeconomics. As a result, the partial cointegrated models will be recognised as more flexible and practical devices for modelling and analysing non-stationary time series data containing structural breaks. For $I(2)$ models, Paruolo and Rahbek (1999) proposed partial analysis while Kurita et al. (2011) introduced a model with deterministic shifts. In future work, it may be of interest to combine those ideas as well.

The rest of this paper consists of five sections. Section 2 introduces partial cointegrated models subject to deterministic breaks and their moving-average representations. Section 3 derives partial quasi likelihood-based tests for cointegrating rank allowing for the breaks, and explores the limit distributions of the test statistics. In this section, a response surface analysis is performed by using simulated distributions and then the results of the analysis are summarised as a set of statistical tables. An empirical illustration of the proposed methodology is provided in Section 5. Finally, Section 6 gives concluding remarks. This study used Ox (Doornik 2013) and PcGive (Doornik and Hendry 2013) to conduct the simulations and the empirical study, respectively.

## 2. Models and Representations

We introduce partial cointegrated vector autoregressive models with deterministic breaks. Section 2.1 reviews the existing models known, while Sections 2.2-2.4 provide details of the proposed models.

### 2.1. Previous Models

The cointegrated vector autoregressive model was proposed by Johansen $(1988,1995)$. Suppose that we observe a $p$-variate vector time series $X_{t}$ integrated of order 1, denoted as $I(1)$ hereafter. In the presence of two lags, a constant and restricted linear trend, the model equation for $X_{t}$

$$
\begin{equation*}
\Delta X_{t}=\left(\Pi, \Pi_{\ell}\right)\binom{X_{t-1}}{t}+\Gamma \Delta X_{t-1}+\mu+\varepsilon_{t} \quad \text { for } \quad t=3, \ldots, T \tag{1}
\end{equation*}
$$

with index $\ell$ for the linear trend model and the associated cointegrating rank hypothesis, for $r \leq p$,

$$
\begin{equation*}
\operatorname{rank}\left(\Pi, \Pi_{\ell}\right) \leq r \quad \text { so that } \quad\left(\Pi, \Pi_{\ell}\right)=\alpha\left(\beta^{\prime}, \gamma\right) \tag{2}
\end{equation*}
$$

Here, the initial values $X_{1}$ and $X_{2}$ are fixed while $p$-vector innovations $\varepsilon_{3}, \ldots, \varepsilon_{T}$ are distributed as independent normal, denoted by $\mathrm{N}_{p}(0, \Omega)$. The parameters in Equation (1) are all variation free, defined as $\alpha, \beta \in \boldsymbol{R}^{p \times r}, \gamma \in \boldsymbol{R}^{r}, \mu \in \boldsymbol{R}^{p}$ and $\Gamma, \Omega \in \boldsymbol{R}^{p \times p}$ and with $\Omega$ being positive definite. This model is interpreted in terms of its Granger-Johansen representation. The likelihood function is maximised through reduced rank regression of $\Delta X_{t}$ on the vector of $X_{t-1}, 1$ corrected for $\Delta X_{t-1}$. The cointegrating rank $r$ can be determined through a sequence of rank test statistics, which have Dickey-Fuller type limit distributions depending on the number of common trends, $p-r$ in this case, and with a linear trend adjustment. Once the rank is determined, asymptotic inference for the cointegrating vectors $\beta$ and the adjustment vectors $\alpha$ can be based on $\chi^{2}$ distributions.

The partial model is derived from the model given by Equation (1), which is referred to as the full model henceforth. This allows exogenous regressors that are not necessarily analysed in the model equation. With a view to setting up the partial model, let us introduce an integer $m$ satisfying $0 \leq r \leq m<p$, so that we can decompose $X_{t}$ into an $m$-vector $Y_{t}$ and a vector $Z_{t}$ of dimension $p-m$. Decompose the parameters and error terms of Equation (1) conformably so that, for instance,

$$
\Pi=\binom{\Pi_{y}}{\Pi_{z}}, \Gamma=\binom{\Gamma_{y}}{\Gamma_{z}}, \mu=\binom{\mu_{y}}{\mu_{z}}, \varepsilon_{t}=\binom{\varepsilon_{y, t}}{\varepsilon_{z, t}} \text { and } \Omega=\left(\begin{array}{ll}
\Omega_{y y} & \Omega_{y z} \\
\Omega_{z y} & \Omega_{z z}
\end{array}\right)
$$

We also define the population regression coefficient $\omega=\Omega_{y z} \Omega_{z z}^{-1}$, which leads to a class of conditional coefficients $\Pi_{y \cdot z}=\Pi_{y}-\omega \Pi_{z}, \Gamma_{y \cdot z}=\Gamma_{y}-\omega \Gamma_{z}$ and $\mu_{y \cdot z}=\mu_{y}-\omega \mu_{z}$. The partial or conditional model for $Y_{t}$ given $Z_{t}$ is then presented as

$$
\begin{equation*}
\Delta Y_{t}=\omega \Delta Z_{t}+\left(\Pi_{y \cdot z}, \Pi_{y \cdot z, \ell}\right)\binom{X_{t-1}}{t}+\Gamma_{y \cdot z} \Delta X_{t-1}+\mu_{y \cdot z}+\varepsilon_{y \cdot z, t} \tag{3}
\end{equation*}
$$

where the conditional innovation sequence $\varepsilon_{y \cdot z, t}=\varepsilon_{y, t}-\omega \varepsilon_{z, t}$ is $\mathrm{N}_{m}\left(0, \Omega_{y y \cdot z}\right)$ distributed, so $\varepsilon_{y \cdot z, t}$ is independent of $Z_{t}$ and the overall past series, while its variance is

$$
\begin{equation*}
\Omega_{y y \cdot z}=\Omega_{y y}-\Omega_{y z} \Omega_{z z}^{-1} \Omega_{z y} . \tag{4}
\end{equation*}
$$

The cointegration rank hypothesis is, for $r \leq m$,

$$
\begin{equation*}
\operatorname{rank}\left(\Pi_{y \cdot z}, \Pi_{y \cdot z, \ell}\right) \leq r \quad \text { so that } \quad\left(\Pi_{y \cdot z}, \Pi_{y \cdot z, \ell}\right)=\alpha_{y \cdot z}\left(\beta^{\prime}, \gamma\right) \tag{5}
\end{equation*}
$$

where $\alpha_{y \cdot z}=\alpha_{y}-\omega \alpha_{z}$. The marginal model for $Z_{t}$ is simply given as

$$
\begin{equation*}
\Delta Z_{t}=\alpha_{z}\left(\beta^{\prime}, \gamma\right)\binom{X_{t-1}}{t}+\Gamma_{z} \Delta X_{t-1}+\mu_{z}+\varepsilon_{z, t} \tag{6}
\end{equation*}
$$

Due to the conditioning of $Y_{t}$ on $Z_{t}$, the innovations $\varepsilon_{y \cdot z, t}$ and $\varepsilon_{z, t}$ are independent. Even so, the cointegrating relationships $\beta^{\prime} X_{t-1}+\gamma t$ form cross equation restrictions, so that maximum likelihood estimation involves a joint analysis of (3) and (6). The rank can be determined from a partial analysis using information criteria albeit without size control as argued by Cavaliere et al. (2018).

Weak exogeneity arises when $\alpha_{z}=0$. In this case, the partial model and the marginal model are unrelated and $Z_{t}$ is weakly exogenous for a class of parameters of interest, $\alpha_{y}, \beta$ and $\gamma$, in the sense of Engle et al. (1983). See also Johansen (1992a, 1992b, 1995, §8) and HJNR. Maximum likelihood estimation can be performed by analysing the two models separately, i.e., the partial model is estimated by reduced rank regression while the marginal model is by least squares regression. The maintained assumption is that the joint vector $X_{t}$ has $r$ cointegrating relations and hence $p-r$ common trends, with the cointegrating relations being in the partial model for $Y_{t}$. A notable feature of the setup is that it is left unspecified whether or not $Z_{t}$ is cointegrated. In a one-lag model $Z_{t}$ will not be cointegrated, but with further lags $Z_{t}$ could be cointegrated since the short-run dynamics are determined by both $\alpha$ and $\Gamma$; see HJNR (p. 390) for an example of these models. HJNR explored an asymptotic theory for likelihood-based rank testing in the partial model (3). The asymptotic distribution of HJNR's rank test statistic is of the Dickey-Fuller type, now depending on both $m-r$ and $p-r$, which are the dimensions of common trends for $Y_{t}$ and $X_{t}$, respectively. Seo (1998) suggested a class of cointegrated models where a stationary regressor, $\Delta Z_{t}$ is included in a cointegration model. This corresponds to a models of the type (3), but where $X_{t-1}$ is replaced by only $Y_{t-1}$. In general, this results in an inference that depends on nuisance parameters. Rahbek and Mosconi (1999) noticed that, if the stationary regressor $\Delta Z_{t}$ is cumulated and entered in the cointegrating vector $X_{t-1}$ as in (3), then the asymptotic distributions of HJNR would apply.

Structural breaks in deterministic terms were included in the full system model by JMN. The idea is to consider, say, two sub-samples starting at time $T_{0}$ and $T_{1}$, respectively, for $0=T_{0}<T_{1}<T_{2}=T$. The dynamic parameters in the model are the same for both sub-samples, while the parameters for deterministic terms can differ. In the model with lag-length $k=2$, the observations $T_{j-1}+1, T_{j-2}+2$ for $j=1,2$ are held back as initial observations. Thus, the transition from one regime to the next is not modelled. Recently, Harvey and Thiele (2017) used a similar idea in a structural time series model.

### 2.2. The Partial Model with Structural Breaks

We are in a position to introduce a new model, a partial cointegrated model allowing for structural breaks in its deterministic terms.

We start by defining the timing of the sub-samples. Suppose we have $T$ observations. We extend the partial model to the one with a pre-specified number of sub-sample periods, $q$ say, and $k$ lags. Following JMN, we introduce the sub-sample structure $0=T_{0}<T_{1}<\cdots<T_{q}=T$. The model will
have $k$ lags. Thus, for each sub-sample $j$, the effective range is $T_{j-1}+k<t \leq T_{j}$. In summary, we have data for $0<t \leq T$, while the effective sample is the collection of effective sub-samples, that is,

$$
\begin{equation*}
T_{j-1}+k<t \leq T_{j} \quad \text { where } \quad 1 \leq j \leq q . \tag{7}
\end{equation*}
$$

The model has dynamic parameters that are common across the sub-sample periods, whereas the parameters for deterministic terms vary. This gives, for each effective sub-sample $j$ as defined above:

$$
\begin{equation*}
\Delta Y_{t}=\omega \Delta Z_{t}+\alpha_{y}\left(\beta^{\prime}, \gamma_{j}\right)\binom{X_{t-1}}{t}+\sum_{i=1}^{k-1} \Gamma_{y \cdot z, i} \Delta X_{t-i}+\mu_{y \cdot z, j}+\varepsilon_{y \cdot z, t} \tag{8}
\end{equation*}
$$

where $\gamma_{j} \in \boldsymbol{R}^{r}, \mu_{j}=\left(\mu_{y, j}^{\prime}, \mu_{z, j}^{\prime}\right)^{\prime} \in \boldsymbol{R}^{p}$ and $\mu_{y \cdot z, j}=\mu_{y, j}-\omega \mu_{z, j}$ for $j=1, \ldots, q$, along with $\Gamma_{i}=\left(\Gamma_{y, i}^{\prime}, \Gamma_{z, i}^{\prime}\right)^{\prime} \in R^{p \times p}$ and $\Gamma_{y-z, i}=\Gamma_{y, i}-\omega \Gamma_{z, i}$ for $i=1, \ldots, k-1$, and all the other parameters were defined in the previous sub-section. Note that the parameters for deterministic terms depend on $j$, indicating the presence of parameter shifts according to regime changes. A class of initial observations $X_{T_{j-1}+1}, \ldots, X_{T_{j-1}+k}$ plays the dual role of capturing the transition from the previous regime, $j-1$ and of serving as the initial observations for the regime $j$. In some applications, the transition between the regimes may be longer than $k$ observations, in which case more observations could be classified as initial observations. The marginal model for $Z_{t}$ under $\alpha_{z}=0$ is

$$
\begin{equation*}
\Delta Z_{t}=\sum_{i=1}^{k-1} \Gamma_{z} \Delta X_{t-i}+\mu_{z, j}+\varepsilon_{z, t} \tag{9}
\end{equation*}
$$

We can form a full model equation as in Equation (1) for each sub-sample period. This is the model of JMN with weak exogeneity imposed. This model will be presented in the next sub-section.

The partial model can be formulated as a single equation for the full sample period in terms of the following notation. Following JMN, we define impulse dummy variables as

$$
D_{j, t}=\left\{\begin{array}{ll}
1 & \text { for } t=T_{j-1}, \\
0 & \text { otherwise, }
\end{array} \quad \text { for } j=1, \ldots, q \text { and } t=1, \ldots, T,\right.
$$

so that $D_{j, t-i}=1$ if $t=T_{j-1}+i$, and also define indicators for the effective samples as

$$
E_{j, t}=\sum_{i=k+1}^{T_{j}-T_{j-1}} D_{j, t-i}=\left\{\begin{array}{ll}
1 & \text { for } T_{j-1}+k<t \leq T_{j}, \\
0 & \text { otherwise },
\end{array} \quad \text { and } \quad E_{t}=\left(E_{1, t}, \ldots, E_{q, t}\right)^{\prime} .\right.
$$

The whole-sample model equation then has the form, with $X_{t-1}^{\ell}=\left(X_{t-1}^{\prime}, t E_{t}^{\prime}\right)^{\prime}$, where the index $\ell$ indicates the model with a linear trend, for $t=k+1, \ldots, T$,

$$
\begin{equation*}
\Delta Y_{t}=\omega \Delta Z_{t}+\alpha_{y}\left(\Pi_{y}, \Pi_{y, \ell}\right) X_{t-1}^{\ell}+\sum_{i=1}^{k-1} \Gamma_{y \cdot z, i} \Delta X_{t-i}+\mu_{y \cdot z} E_{t}+\sum_{i=1}^{k} \sum_{j=2}^{q} \varphi_{j, i} D_{j, t-i}+\varepsilon_{y \cdot z, t} \tag{10}
\end{equation*}
$$

with cointegration rank hypothesis, for $r \leq m$,

$$
\begin{equation*}
\mathrm{H}_{\ell}(r): \quad \operatorname{rank}\left(\Pi_{y}, \Pi_{y, \ell}\right) \leq r \quad \text { so that } \quad\left(\Pi_{y}, \Pi_{y, \ell}\right)=\left(\beta^{\prime}, \gamma\right) \tag{11}
\end{equation*}
$$

Here, $\varphi_{j, i} \in \boldsymbol{R}^{m}$ represents a class of parameters for $D_{j, t-i}$ for $i=1, \ldots, k$ and $j=2, \ldots, q$, while the parameters $\gamma$ and $\mu_{y \cdot z}$ are now redefined in a manner allowing for breaks as

$$
\gamma=\left(\gamma_{1}, \ldots, \gamma_{q}\right) \in R^{r \times q} \quad \text { and } \quad \mu_{y \cdot z}=\left(\mu_{y \cdot z, 1}, \ldots, \mu_{y \cdot z, q}\right) \in R^{m \times q}
$$

which are used in the rest of this study. Equation (8), or its whole-sample form (10), is referred to as the partial model with a broken linear trend term.

### 2.3. Representations

Various properties of the proposed partial model (8) will be analysed using the Granger-Johansen representation of an $I(1)$ process, which is formulated based on the full model for $X_{t}$; thus, the representation is the same as that in JMN (Theorem 2.1). In JMN, each sub-sample period is analysed conditionally on its initial observations. As a result, the representation for each sub-sample period is the same as that in Johansen (1995, Theorem 4.2). The initial values for each sub-sample can be large and thus be influential even in the asymptotic context, but, when following the underlying argument of JMN, one can see that such initial values do not play critical roles in the required asymptotic analysis. Following Kurita and Nielsen (2009), we show this in two steps: first, we analyse a homogeneous equation, and then consider the roles of deterministic terms by moving to a non-homogeneous equation. For further details, see the proof of Theorem 1 below.

For each sub-sample, the full model for $X_{t}$ is defined as a joint system of (8) and (9) through:

$$
\begin{equation*}
\Delta X_{t}=\alpha\left(\beta^{\prime}, \gamma_{j}\right)\binom{X_{t-1}}{t}+\sum_{i=1}^{k-1} \Gamma_{i} \Delta X_{t-i}+\mu_{j}+\varepsilon_{t} \tag{12}
\end{equation*}
$$

while the corresponding homogeneous equation is

$$
\begin{equation*}
\Delta \tilde{X}_{t}=\alpha \beta^{\prime} \tilde{X}_{t-1}+\sum_{i=1}^{k-1} \Gamma_{i} \Delta \tilde{X}_{t-i}+\varepsilon_{t} \tag{13}
\end{equation*}
$$

where $\tilde{X}_{t}$ denotes a $p$-variate mean-zero vector time series. We then set up a companion vector based on (13) and analyse a companion form of this equation. Several choices are conceivable with respect to a companion form for (13) and we use the choice that appears, for instance, in Hansen (2005). For the purpose of studying details of the representation, the parameters need to satisfy Assumption 1 below. This is applicable to both (12) and (13). Some additional notation is required. When $\beta$ has full column rank $r$, let $\beta_{\perp}$ denote a $p \times(p-r)$ dimensional orthogonal complement, so that $\left(\beta, \beta_{\perp}\right)$ is invertible and $\beta_{\perp}^{\prime} \beta=0$ and introduce the normalization $\bar{\beta}=\beta\left(\beta^{\prime} \beta\right)^{-1}$. The same notation applies to $\alpha$.

Assumption 1. Assume that the roots of the characteristic polynomial,

$$
A(z)=(1-z) I_{p}-\alpha \beta^{\prime} z-\sum_{i=1}^{k-1} \Gamma_{i}(1-z) z^{i}
$$

are outside the complex unit circle or at unity; furthermore, assume that the matrices $\alpha$ and $\beta$ have full column rank $r$ and that the square matrix $\alpha_{\perp}^{\prime} \Psi \beta_{\perp}$ has full rank $p-r$, where $\Psi=I_{p}-\sum_{i=1}^{k-1} \Gamma_{i}$.

Given Assumption 1, we can define $C=\beta_{\perp}\left(\alpha_{\perp}^{\prime} \Psi \beta_{\perp}\right)^{-1} \alpha_{\perp}^{\prime}$ This is often referred to as the impact matrix in cointegration literature; see Paruolo (1997) for inference on this matrix.

We are approaching the stage where the Granger-Johansen representation for each sub-sample period is presented. For the homogeneous Equation (13), let us define

$$
\alpha=\left(\begin{array}{cccc}
\alpha & \Gamma_{1} & \cdots & \Gamma_{k-1}  \tag{14}\\
0 & I_{p} & & 0 \\
\vdots & & \ddots & \vdots \\
& & & 0 \\
0 & \cdots & & 0
\end{array} I_{p} .\right), \quad \Lambda=\left(\begin{array}{ccccc}
I_{p} & 0 & \cdots & & 0 \\
I_{p} & -I_{p} & & \vdots \\
0 & & \ddots & & \\
\vdots & & & & 0 \\
0 & \cdots & 0 & I_{p} & -I_{p}
\end{array}\right)
$$

as well as

$$
\beta^{\prime}=\left(\begin{array}{cc}
\beta^{\prime} & 0  \tag{15}\\
0 & I_{p(k-1)}
\end{array}\right) \Lambda, \quad \tilde{X}_{t-1}=\left(\begin{array}{c}
X_{t-1} \\
\vdots \\
X_{t-k}
\end{array}\right)
$$

and $\iota=\left(I_{p}, 0, \ldots, 0\right)^{\prime}$ together with $r=r+p(k-1)$. The representation is then given in the theorem below, the proof of which is provided in Appendix B.

Theorem 1. Suppose that Assumption 1 is fulfilled. Then, an $\boldsymbol{r}$-variate process $\boldsymbol{\beta}^{\prime} \tilde{\boldsymbol{X}}_{t}$ derived from ation (13) satisfies, on the effective sample $T_{j-1}+k<t \leq T_{j}$ for $1 \leq j \leq q$,

$$
\begin{equation*}
\boldsymbol{\beta}^{\prime} \tilde{X}_{t}=\left(I_{r}+\boldsymbol{\beta}^{\prime} \boldsymbol{\alpha}\right) \boldsymbol{\beta}^{\prime} \tilde{X}_{t-1}+\boldsymbol{\beta}^{\prime} \boldsymbol{\iota} \varepsilon_{t} \quad \text { with } \quad\left|\operatorname{eigen}\left(I_{r}+\boldsymbol{\beta}^{\prime} \boldsymbol{\alpha}\right)\right|<1 \tag{16}
\end{equation*}
$$

which is a stable first-order vector autoregression. The solution to (13) is given as

$$
\begin{equation*}
\tilde{X}_{t}=C \sum_{s=T_{j-1}+k+1}^{t} \varepsilon_{s}+\{(I-C \Psi) \bar{\beta}, C Y\} \boldsymbol{\beta}^{\prime} \tilde{\boldsymbol{X}}_{t}-C(\Psi, \mathrm{Y}) \Lambda \tilde{\boldsymbol{X}}_{T_{j-1}+k} \tag{17}
\end{equation*}
$$

where $\mathrm{Y}=\left(\mathrm{Y}_{1}, \ldots \mathrm{Y}_{k-1}\right)$ with $\mathrm{Y}_{i}=-\Gamma_{i}-\cdots-\Gamma_{k-1}$. Thus, the variable $X_{t}$ in (12) satisfies

$$
\begin{align*}
X_{t}=C \sum_{s=T_{j-1}+k+1}^{t} \varepsilon_{s}+(I-C \Psi) \bar{\beta} \beta^{\prime} \tilde{X}_{t}-C & \sum_{i=1}^{k-1} \Gamma_{i} \sum_{\ell=0}^{i-1} \Delta \tilde{X}_{t-\ell} \\
& -C \Psi \tilde{X}_{T_{j-1}+k}+C \sum_{i=1}^{k-1} \Gamma_{i} \sum_{\ell=0}^{i-1} \Delta \tilde{X}_{T_{j-1}+k-\ell}+\tau_{c, j}+\tau_{\ell, j} t \tag{18}
\end{align*}
$$

for $\tilde{X}_{t}=X_{t}-\tau_{c, j}-\tau_{\ell, j} t$ with the parameters $\tau_{c, j}$ and $\tau_{\ell, j}$ satisfying

$$
\Psi \tau_{l, j}=\alpha \beta^{\prime}\left(\tau_{c, j}-\tau_{\ell, j}\right)+\mu_{j} \quad \text { and } \quad \beta^{\prime} \tau_{\ell, j}+\gamma_{j}=0 .
$$

Note that the initial observations for the $j$-th sub-sample in (18) are expressed in terms of linear combinations of the mean-zero values $\tilde{X}_{T_{j-1}+1}, \ldots, \tilde{X}_{T_{j-1}+k}$, so that we can in general argue that the the starting values for each sub-sample period do not play critical roles in asymptotic analysis. This property was not explicitly examined in JMN. Thus, Theorem 1 can be seen as a useful clarification of roles of the initial values in the full cointegrated model subject to deterministic breaks. The Granger-Johansen representation is utilised in proofs of asymptotic theorems in Section 3.

As an alternative to the above sub-sample representation, one can derive a joint representation for the whole sample. For this purpose, we need a full system equation for $X_{t}$ over the entire sample period. This equation is derived from a combination of (12) over $j=1, \ldots, q$ augmented with dummies $D_{j, t-i}$ and $E_{j, t}$, as in (10); that is,

$$
\begin{equation*}
\Delta X_{t}=\alpha\left(\beta^{\prime}, \gamma\right) X_{t-1}^{\ell}+\sum_{i=1}^{k-1} \Gamma_{i} \Delta X_{t-i}+\mu E_{t}+\sum_{i=1}^{k} \sum_{j=2}^{q} \kappa_{j, i} D_{j, t-i}+\varepsilon_{t} \tag{19}
\end{equation*}
$$

where $\kappa_{j, i} \in R^{p}$ for $i=1, \ldots, k$ and $j=2, \ldots, q$, and $\mu=\left(\mu_{1}, \ldots, \mu_{q}\right) \in \boldsymbol{R}^{p \times q}$; see Equation (2.6) in JMN . We then replace the innovations $\varepsilon_{t}$ with $\varepsilon_{t}^{D}=\varepsilon_{t}+\alpha \gamma t E_{t}+\mu E_{t}+\sum_{i=1}^{k} \sum_{j=2}^{q} \kappa_{j, i} D_{j, t-i}$ to reach a whole-sample representation such as

$$
\begin{equation*}
X_{t} \approx C \sum_{s=k+1}^{t} \varepsilon_{s}^{D}+C_{1}(L) \varepsilon_{t}^{D}+A \quad \text { for } \quad k<t \leq T \tag{20}
\end{equation*}
$$

where $C_{1}(L) \varepsilon_{t}^{D}$ denotes a moving-average process whose coefficients decrease exponentially fast, and $A$ depends on initial observations $X_{1}, \ldots, X_{k}$, satisfying $\beta^{\prime} A=0$. This is an approximation, since the precise formulation of the moving-average component requires introduction of an infinite past, while the model is formulated as conditional on the initial observations. As before the deterministic parts of the common trends $C \sum_{s=k+1}^{T} \varepsilon_{s}^{D}$ will be piecewise constant since $C \alpha=0$, so that each constant fails to cumulate to a linear trend. A similar approach was adopted in $I(2)$ cointegration analysis by Kurita et al. (2011). The representation (20) is clear and concise, but the transition from one regime to another is considered to be explicitly autoregressive, which may leave less flexibility to represent regime transitions of some persistent and messy nature. For the asymptotic study conducted below, we follow JMN by using the sub-sample representation (18).

### 2.4. The Partial Model with Shifts in The Level

In some applications, it suffices to exclude the broken linear trends and just include shifts in the constant term. By restricting the broken constant term within the cointegrating space, Equation (10) is reduced to, with $X_{t-1}^{c}=\left(X_{t-1}^{\prime}, E_{t}^{\prime}\right)^{\prime}$ and index $c$ for model with breaks in the constant level,

$$
\begin{equation*}
\Delta Y_{t}=\omega \Delta Z_{t}+\left(\Pi_{y}, \Pi_{y, c}\right) X_{t-1}^{c}+\sum_{i=1}^{k-1} \Gamma_{y \cdot z, i} \Delta X_{t-i}+\sum_{i=1}^{k} \sum_{j=2}^{q} \varphi_{j, i} D_{j, t-i}+\varepsilon_{y \cdot z, t} \tag{21}
\end{equation*}
$$

and cointegration rank hypothesis, for $r \leq m$,

$$
\begin{equation*}
\mathrm{H}_{c}(r): \quad \operatorname{rank}\left(\Pi_{y}, \Pi_{y, c}\right) \leq r \quad \text { or that } \quad\left(\Pi_{y}, \Pi_{y, c}\right)=\alpha_{y}\left(\beta^{\prime}, \gamma\right) \tag{22}
\end{equation*}
$$

The Granger-Johansen representation has the same form as (18) but is subject to

$$
\beta^{\prime} \tau_{c, j}+\gamma_{j}^{\prime}=0 \quad \text { and } \quad \tau_{\ell, j}=0
$$

## 3. Testing for Cointegrating Rank in the Partial Models

This section addresses the issue of testing for cointegrating rank in the suggested partial models with deterministic shifts. Section 3.1 introduces a partial likelihood ratio test for the choice of rank based on the broken linear-trend model and Section 3.2 derives its limit distribution. Section 3.3 then turns to the broken constant model and examines the test statistic based upon it. Finally, Section 4 derives a class of approximations to the limit distributions by means of computer simulations and response surface regression.

### 3.1. Rank Test Statistic

For each sub-sample period, the partial model (10) is seen as equivalent to that in HJNR, given the presence of structural breaks in its deterministic terms. We derive the log partial likelihood ratio test statistic for the cointegration rank hypothesis $\mathrm{H}_{\ell}(r)$ defined in (11). This likelihood is analysed by reduced rank regression in a manner similar to the original cointegration model in Johansen $(1988,1995)$. We show that the reduced rank regression can be done in three, numerically equivalent ways.

The first approach is based on a full-sample reduced rank regression. Regress each of the vectors $\Delta Y_{t}$ and $X_{t-1}^{\ell}=\left(X_{t-1}^{\prime}, t E_{t}^{\prime}\right)^{\prime}$ on a vector $H_{t}$ consisting of the variables $\Delta Z_{t}$, the lagged differences $\Delta X_{t-1}, \ldots, \Delta X_{t-k+1}$, the intercepts $E_{t}$ and the impulse dummies $D_{j, t-i}$ for $i=1, \ldots, k$ and $j=2, \ldots, q$, so that $H_{t}$ has dimension $p k-m+q+k(q-1)$. This gives residuals $R_{0, t}$ and $R_{1, t}$ :

$$
\binom{R_{0, t}}{R_{1, t}}=\binom{\Delta Y_{t}}{X_{t-1}^{\ell}}-\sum_{s=k+1}^{T}\binom{\Delta Y_{s}}{X_{s-1}^{\ell}} H_{s}^{\prime}\left(\sum_{s=k+1}^{T} H_{s} H_{s}^{\prime}\right)^{-1} H_{t} \quad \text { for } k<t \leq T .
$$

The second approach is viewed as a sub-sample approach. We note that the impulse dummies result in a perfect fit for each transitional period in between two connecting regimes; thus, $R_{0, t}$ and $R_{1, t}$ are zero for all transitional periods. We can, therefore, compute the residuals $R_{0, t}$ and $R_{1, t}$ by analysing the effective sub-sample periods only; see also Doornik et al. $(1998, \S 12.2)$. For this purpose, let us form regressors $P_{t}$ from the variables $\Delta Z_{t}$, the lagged differences $\Delta X_{t-1}, \ldots, \Delta X_{t-k+1}$ and the intercepts $E_{t}$, so that $P_{t}$ is a vector of dimension $p k-m+q$. The residuals $R_{0, t}$ and $R_{1, t}$ then satisfy

$$
\binom{R_{0, t}}{R_{1, t}}=\binom{\Delta Y_{t}}{X_{t-1}^{\ell}}-\sum_{j=1}^{q} \sum_{s=T_{j-1}+k+1}^{T_{j}}\binom{\Delta Y_{s}}{X_{s-1}^{\ell}} P_{s}^{\prime}\left(\sum_{j=1}^{q} \sum_{s=T_{j-1}+k+1}^{T_{j}} P_{s} P_{s}^{\prime}\right)^{-1} P_{t}
$$

for $T_{j-1}+k<t \leq T_{j}$ with $1 \leq j \leq q$, while $R_{0, t}$ and $R_{1, t}$ are zero, otherwise.
The third approach is recognised as a two-step approach, in which we first demean the observed time series and then partial out influences from the lagged differences. In the first step, we analyse two vectors $\Delta Y_{t}$ and $X_{t-1}^{\ell}$, along with a vector $V_{t}$ consisting of the variables $\Delta Z_{t}$ and the lagged differences $\Delta X_{t-1}, \ldots, \Delta X_{t-k+1}$. These three vectors are demeaned within each sub-sample period, yielding $\mathcal{Z}_{0, t}, \mathcal{Z}_{1, t}$ and $\mathcal{Z}_{2, t}$ defined as

$$
\left(\begin{array}{c}
\mathcal{Z}_{0, t}  \tag{23}\\
\mathcal{Z}_{1, t} \\
\mathcal{Z}_{2, t}
\end{array}\right)=\left(\begin{array}{c}
\Delta Y_{t} \\
X_{t-1}^{\ell} \\
V_{t}
\end{array}\right)-\frac{1}{T_{j}-T_{j-1}+k} \sum_{s=T_{j-1}+k+1}^{T_{j}}\left(\begin{array}{c}
\Delta Y_{s} \\
X_{s-1}^{\ell} \\
V_{s}
\end{array}\right)
$$

for $1 \leq j \leq q$ and $T_{j-1}+k<t \leq T_{j}$ and zero otherwise. In the second step, we compute

$$
\binom{R_{0, t}}{R_{1, t}}=\binom{\mathcal{Z}_{0, t}}{\mathcal{Z}_{1, t}}-\sum_{s=k+1}^{T}\binom{\mathcal{Z}_{0, s}}{\mathcal{Z}_{1, s}} \mathcal{Z}_{2, s}^{\prime}\left(\sum_{s=k+1}^{T} \mathcal{Z}_{2, s} \mathcal{Z}_{2, s}^{\prime}\right)^{-1} \mathcal{Z}_{2, t}
$$

Since $\mathcal{Z}_{0, t}, \mathcal{Z}_{1, t}$ are zero within the transitional periods, so are the residuals $R_{0, t}, R_{1, t}$.
With the residuals $R_{0, t}$ and $R_{1, t}$ in hand, we can compute the product moments

$$
\left(\begin{array}{ll}
S_{00} & S_{01}  \tag{24}\\
S_{10} & S_{11}
\end{array}\right)=\frac{1}{T-k} \sum_{t=k+1}^{T}\binom{R_{0, t}}{R_{1, t}}\binom{R_{0, t}}{R_{1, t}}^{\prime}
$$

and a set of squared canonical correlations $1 \geq \hat{\lambda}_{1} \geq \cdots \geq \hat{\lambda}_{m} \geq 0$ by solving the eigenvalue problem

$$
0=\operatorname{det}\left(\lambda S_{11}-S_{10} S_{00}^{-1} S_{01}\right)
$$

Hence, the log partial likelihood ratio $(P L R)$ test statistic for the null hypothesis of cointegrating rank $r, \mathrm{H}_{\ell}(r)$, against the hypothesis $\mathrm{H}_{\ell}(m)$ is

$$
\begin{equation*}
\operatorname{PLR}\left\{\mathrm{H}_{\ell}(r) \mid \mathrm{H}_{\ell}(m)\right\}=-(T-k) \sum_{i=r+1}^{m} \log \left(1-\hat{\lambda}_{i}\right) \tag{25}
\end{equation*}
$$

### 3.2. Asymptotic Distribution of the Test Statistic

We derive the asymptotic distribution of the rank test statistic in a setting where the relative break points satisfy $T_{j} / T \rightarrow v_{j}$ for $j=0, \ldots, q$ while $T$ goes to infinity. The relative break points satisfy $0=v_{0}<v_{1}<\cdots<v_{q}=1$. Indeed, with $\operatorname{int}(x)$ denoting the integer part of $x$, then the $q$-vector $E_{\text {int }(T u)}$ on $u \in[0,1]$ has the limit

$$
\begin{equation*}
e_{u}=\left\{1_{\left(v_{0}<u<v_{1}\right)}, \ldots, 1_{\left(v_{q-1}<u<v_{q}\right)}\right\}^{\prime} \tag{26}
\end{equation*}
$$

In the standard framework developed by Johansen (1995), the innovation sequence $\varepsilon_{t}$ is assumed to be independent and identically Gaussian distributed, and this assumption was adopted by HJNR and JMN as reviewed in Section 2.1 above. We relax this normality assumption to a martingale difference assumption. If the innovations $\varepsilon_{t}$ are not normal, the model equations lead to a quasi-likelihood function rather than a likelihood function. Weak exogeneity is preserved as it is a property of the likelihood rather than the distribution of the innovations as such. The partial innovation $\varepsilon_{y \cdot z, t}=\varepsilon_{y, t}-\omega \varepsilon_{z, t}$ and the marginal innovation $\varepsilon_{z, t}$ are uncorrelated, but they will not be independent in general when moving away from the normality assumption. We can no longer appeal to the conditional-distribution argument, as implied in Equation (3). Thus, the conditional-distribution argument is replaced with a regression argument, see Appendix C.2. The martingale difference assumption is summarised as Assumption 2 below.

Assumption 2. Assume that $\varepsilon_{t}$ is a martingale difference sequence with respect to a filtration $\mathcal{F}_{t}$ such that $\mathrm{E}\left(\varepsilon_{t} \mid \mathcal{F}_{t-1}\right)=0$ almost surely (a.s.). Let $\Omega$ be a positive definite matrix. Suppose that
(i) $\mathrm{E}\left(\varepsilon_{t} \varepsilon_{t}^{\prime}\right)=\Omega$;
(ii) $T^{-1} \sum_{t=1}^{T} \mathrm{E}\left(\varepsilon_{t} \varepsilon_{t}^{\prime} \mid \mathcal{F}_{t-1}\right) \xrightarrow{\mathrm{P}} \Omega$;
(iii) either of the following boundedness conditions
(a) $\sup _{t \in \mathbb{N}} \mathrm{E}\left\{\varepsilon_{t}^{\prime} \varepsilon_{t} 1_{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>a\right)} \mid \mathcal{F}_{t-1}\right\} \xrightarrow{\mathrm{P}} 0$ as $a \rightarrow \infty$;
(b) $\sup _{t \in \mathbb{N}} \mathrm{E}\left|\varepsilon_{t}\right|^{4}<\infty$.

The boundedness conditions in part (iii) are not nested. Part (a) can be satisfied without the existence of fourth moments as in part (b). Conversely, bounded fourth moments in part $(b)$ do not necessarily imply part (a); see Remark A1 in Appendix C.1.

Under Assumption 2, we are able to apply the results of Brown (1971) to analyse the random walk components of the process. For this, we require a Lindeberg condition, which is established in Lemma A1 in Appendix C. 1 under Assumption 2. Brown's result is for univariate martingale difference sequences and requires that the ratio of the sum of conditional variances to that of unconditional variances should converge to unity. For the multivariate case, we can apply the Cramér-Wold device and form linear combinations of the present multivariate martingale differences. Using parts (i), (ii) we can then show that Brown's ratios converge to unity.

Under Assumption 2, we can also analyse the (approximately) stationary components of the process. Under part (iii.a), we can apply the results of Anderson and Kunitomo (1992), which exploit a truncation argument. Under part (iii.b) we can apply the same ideas as in Anderson and Kunitomo (1992) but without the truncation argument.

Cointegration models with heteroscedasticity have previously been analysed by for instance Cavaliere et al. (2010) and Boswijk et al. (2016). The former paper is concerned with rank testing in a full system. For the analysis of the (approximately) stationary components, it relies on Hannan and Heyde (1972) who require an almost sure version of Assumption 2(ii). The latter paper is concerned with testing on the cointegrating vectors in a full system with an elaborate, deterministic structure for the variances of the innovations.

Before proceeding to the main results, we present a set of stronger assumptions requiring constant conditional variance, see Assumption 3 and Lemma 1 below. These assumptions were used by Lai and Wei $(1982,1985)$ as well as Chan and Wei $(1988)$. They have two advantages. First, they are easier to check for practitioners than the convergence results in Assumption 2. Second, the assumptions could be used to derive a variety of almost sure convergence results, as explored by Lai and Wei $(1982,1985)$ and Nielsen $(2005)$, although we will not exploit those properties here.

Assumption 3. Assume that $\varepsilon_{t}$ is a martingale difference sequence with respect to a filtration $\mathcal{F}_{t}$ such that $\mathrm{E}\left(\varepsilon_{t} \mid \mathcal{F}_{t-1}\right)=0$ a.s. and
(i) $\operatorname{Var}\left(\varepsilon_{t} \mid \mathcal{F}_{t-1}\right)=\Omega$ a.s., where $\Omega$ is positive definite;
(ii) $\sup _{t \in \mathbb{N}} \mathrm{E}\left(\left|\varepsilon_{t}\right|^{2+\xi} \mid \mathcal{F}_{t-1}\right)<\infty$ a.s. for some $\xi>0$.

Lemma 1. Assumption 3 implies Assumption 2.
We can now present the limit distribution of the PLR statistic (25), noting that, formally, it is a a log partial quasi-likelihood ratio test statistic under Assumption 3. For this purpose, let $\xrightarrow{\mathrm{D}}$ signify weak convergence, while let $B_{u}$ represent a $(p-r)$-dimensional standard Brownian motion process on $u \in[0,1]$ and let $B_{u}^{(m-r)}$ be the first $m-r$ coordinates of $B_{u}$. The limit distribution is given in the next theorem, which is proved in Appendix C.3.

Theorem 2. Suppose that Assumptions 1 and 2 are satisfied along with $\alpha_{z}=0$, so that $Z_{t}$ is weakly exogenous with respect to $\alpha_{y}, \beta$ and $\gamma$. As $T \rightarrow \infty$, with relative break points satisfying $T_{j} / T \rightarrow v_{j}$ for $0=v_{0}<v_{1}<$ $\cdots<v_{q}=1$, the PLR test statistic (25) under $\mathrm{H}_{\ell}(r)$ satisfies

$$
\begin{equation*}
\operatorname{PLR}\left\{\mathrm{H}_{\ell}(r) \mid \mathrm{H}_{\ell}(m)\right\} \xrightarrow{\mathrm{D}} \mathrm{DF}_{\ell}(m-r, p-r ; v), \tag{27}
\end{equation*}
$$

where, with $e_{u}$ defined in (26),

$$
\begin{aligned}
\operatorname{DF}_{\ell}(m-r, p-r ; v) & =\operatorname{tr}\left\{\int_{0}^{1} d B_{u}^{(m-r)} G_{u}^{\prime}\left(\int_{0}^{1} G_{u} G_{u}^{\prime} d u\right)^{-1} \int_{0}^{1} G_{u} d B_{u}^{(m-r) \prime}\right\} \\
G_{u} & =\binom{B_{u}}{u e_{u}}-\int_{0}^{1}\binom{B_{s}}{s e_{s}} e_{s}^{\prime} d s\left(\int_{0}^{1} e_{s} e_{s}^{\prime} d s\right)^{-1} e_{u}
\end{aligned}
$$

Note that, when $p=m$, the result in Theorem 2 corresponds to Theorem 3.1 in JMN. A direct simulation of (27) is rather laborious. By exploiting some analytic properties of the distributions, we are able to simplify this simulation task. The next Theorem 3 describes these properties by linking the moments of the limit distribution in Theorem 2 to those for the full model. Theorem 3 provides a basis for simulation in Section 4. The proof of this theorem, given in Appendix C.3, is based on a slight modification of results in Doornik (1998, §9); see also Boswijk and Doornik (2005).

Theorem 3. Let $B_{i, u}$ be the $i$-th coordinate of the Brownian motion $B_{u}$. Let

$$
\mathrm{T}_{i}=\int_{0}^{1} d B_{i, u} G_{u}^{\prime}\left(\int_{0}^{1} G_{u} G_{u}^{\prime} d u\right)^{-1} \int_{0}^{1} G_{u} d B_{i, u}^{\prime} \quad \text { for } \quad i=1, \ldots, p-r
$$

Then, $\mathrm{T}_{1}, \ldots, \mathrm{~T}_{p-r}$ are identically distributed and any pairs $\mathrm{T}_{j}, \mathrm{~T}_{k}$ are also identically distributed. Moreover, the limiting statistic (26) satisfies $\mathrm{DF}_{\ell}(m-r, p-r ; v)=\sum_{i=1}^{m-r} \mathrm{~T}_{i}$ with expectation and variance given by

$$
\begin{aligned}
\mathrm{E}\left\{\mathrm{DF}_{\ell}(m-r, p-r ; v)\right\} & =\left(\frac{m-r}{p-r}\right) \mathrm{E}\left(\sum_{i=1}^{p-r} \mathrm{~T}_{i}\right), \\
\operatorname{Var}\left\{\mathrm{DF}_{\ell}(m-r, p-r ; v)\right\} & =\left(\frac{m-r}{p-r}\right) \operatorname{Var}\left(\sum_{i=1}^{p-r} \mathrm{~T}_{i}\right)-(m-r)(p-m) \operatorname{Cov}\left(\mathrm{T}_{1}, \mathrm{~T}_{2}\right) .
\end{aligned}
$$

Since the above Theorem 3 links the moments of the statistics for partial systems and for a full system, we can now proceed by simulating distributions for full systems only. JMN simulated response surfaces for the mean and variance of $\sum_{i=1}^{p-r} \mathrm{~T}_{i}$. As we will also need a response surface for $\operatorname{Cov}\left(\mathrm{T}_{1}, \mathrm{~T}_{2}\right)$, we have to redo their simulation exercise. For this purpose, we quote a result from JMN.

Theorem 4 (JMN, Theorem 3.2). Let $B^{[1]}, \ldots, B^{[q]}$ be independent $(p-r)$-dimensional standard Brownian motions and define

$$
\begin{aligned}
J_{j} & =\left\{\int_{0}^{1}(u \mid 1)^{2} d u\right\}^{-1 / 2} \int_{0}^{1}(u \mid 1)\left\{d B_{u}^{[j](m-r)}\right\} \\
K_{j} & =\int_{0}^{1}\left(B_{u}^{[j]} \mid u, 1\right)\left\{d B_{u}^{[j](m-r)}\right\}^{\prime} \\
L_{j} & =\int_{0}^{1}\left(B_{u}^{[j]} \mid u, 1\right)\left(B_{u}^{[j]} \mid u, 1\right)^{\prime} d u
\end{aligned}
$$

Then, the limiting variable (26) for a full sample with $p=m$ satisfies

$$
\operatorname{DF}_{\ell}(p-r, p-r ; v)=\operatorname{tr}\left[\left(\sum_{j=1}^{q} K_{j} \Delta v_{j}\right)^{\prime}\left\{\sum_{j=1}^{q} L_{j}\left(\Delta v_{j}\right)^{2}\right\}^{-1}\left(\sum_{j=1}^{q} K_{j} \Delta v_{j}\right)\right]+\sum_{j=1}^{q} J_{j}^{\prime} J_{j},
$$

where $\Delta v_{j}=v_{j}-v_{j-1}$. Here, the two summands are independent and $\sum_{j=1}^{q} J_{j}^{\prime} J_{j}$ is distributed as $\chi^{2}\{q(m-r)\}$. Moreover, let $J_{j}^{[i]}$ and $K_{j}^{[i]}$ denote the ith coordinate of $J_{j}$ and $K_{j}$ so that

$$
\mathrm{T}_{i}=\left(\sum_{j=1}^{q} K_{j}^{[i]} \Delta v_{j}\right)^{\prime}\left\{\sum_{j=1}^{q} L_{j}\left(\Delta v_{j}\right)^{2}\right\}^{-1}\left(\sum_{j=1}^{q} K_{j}^{[i]} \Delta v_{j}\right)+\sum_{j=1}^{q}\left(J_{j}^{[i]}\right)^{2} .
$$

As in JMN, we note that Theorem 4 implies a simple relation between the limiting statistics for models with $q$ and with $q-1$ sub-sample periods that is

$$
\begin{equation*}
\lim _{\Delta v_{q} \rightarrow 0} \mathrm{DF}_{\ell}\left(p-r, p-r ; v_{1}, \ldots v_{q-1}, v_{q}\right)=\mathrm{DF}_{\ell}\left(p-r, p-r ; v_{1}, \ldots v_{q-1}\right)+J_{q}^{\prime} J_{q}, \tag{28}
\end{equation*}
$$

where $\mathrm{DF}_{\ell}\left(p-r, p-r ; v_{1}, \ldots v_{q-1}\right)$ and $J_{q}^{\prime} J_{q}$ are independent and $J_{q}^{\prime} J_{q}$ is $\chi^{2}(p-r)$.

### 3.3. Asymptotic Distribution for the Broken Constant Case

The model investigated previously has a broken linear trend. A variant of this model is free of such a linear trend but with a broken constant; see Equation (21). Equation (8) then reduces to

$$
\Delta Y_{t}=\omega \Delta Z_{t}+\alpha_{y}\left(\beta^{\prime}, \gamma_{j}\right)\binom{X_{t-1}}{1}+\sum_{i=1}^{k-1} \Gamma_{y \cdot z, i} \Delta X_{t-i}+\varepsilon_{y \cdot z, t} .
$$

As before, the partial quasi-likelihood is maximized by reduced rank regression. We follow the third approach (23) in Section 3.1, in which the broken linear trend is now replaced with the broken constant, so that we consider the vectors $\Delta Y_{t}$ and $X_{t-1}^{\ell}=\left(X_{t-1}^{\prime}, E_{t}^{\prime}\right)^{\prime}$, together with the vector $V_{t}$ composed of the variables $\Delta Z_{t}$ and the lagged differences $\Delta X_{t-1}, \ldots, \Delta X_{t-k+1}$. Equation (23) then reduces to

$$
\left(\begin{array}{c}
\mathcal{Z}_{0, t} \\
\mathcal{Z}_{1, t} \\
\mathcal{Z}_{2, t}
\end{array}\right)=\left(\begin{array}{c}
\Delta Y_{t} \\
X_{t-1}^{\ell} \\
V_{t}
\end{array}\right)
$$

The limit distribution of the log partial quasi-likelihood ratio test statistic for cointegrating rank $r$, denoted by $\operatorname{PLR}\left\{\mathrm{H}_{c}(r) \mid \mathrm{H}_{c}(m)\right\}$, is given in Theorem 5 below. Its proof is based on a set of modifications of the proofs for the limit theorems in the previous sub-sections.

Theorem 5. Suppose that Assumptions 1 and 2 are satisfied along with $\alpha_{z}=0$, so that $Z_{t}$ is weakly exogenous with respect to $\alpha_{y}, \beta$ and $\gamma$. As $T \rightarrow \infty$, with relative break points satisfying $T_{j} / T \rightarrow v_{j}$ for $0=v_{0}<v_{1}<$ $\cdots<v_{q}=1$, the PLR test statistic (25) under $\mathrm{H}_{c}(r)$ satisfies

$$
\operatorname{PLR}\left\{\mathrm{H}_{c}(r) \mid \mathrm{H}_{c}(m)\right\} \xrightarrow{\mathrm{D}} \mathrm{DF}_{c}(m-r, p-r ; v),
$$

where $\mathrm{DF}_{c}$ is defined as in Theorem 2 with the difference that

$$
G_{u}=\binom{B_{u}}{e_{u}}
$$

The results in Theorems 3 and 4 also apply with the present choice of $G_{u}$.

## 4. Approximations of the Asymptotic Distributions

The limit distributions of the cointegrating rank test statistics are non-standard, as shown in the previous sub-sections; however, given the existing results in the literature, the distributions can be closely approximated by a gamma distribution identified by the first two moments. We first derive this approximation and then show how to implement the approximation.

### 4.1. Derivation of Response Surface

The literature shows that the asymptotic distributions for cointegration rank testing are nearly gamma distributed. The approximating gamma distribution can be captured either through the mean and variance of the asymptotic distribution or through the associated shape and scale parameters. The quality of the gamma-distribution approximation method has been documented in several papers. Using analytic methodology, Nielsen (1997) showed a very good agreement between limit distributions and approximate gamma distributions in tests for unit roots. Doornik (1998) then conducted detailed simulation studies to demonstrate a similar agreement for standard full-system cointegration rank test statistics; see also Doornik (2003) for various tables of asymptotic quantiles produced by the gamma-distribution approximations. JMN also employed this method.

In order to apply the gamma approximation method, we first define parameters for shape and scale. By Theorem 3, the partial system statistic satisfies $\mathrm{DF}_{\ell}(m-r, p-r ; v)=\sum_{i=1}^{m-r} \mathrm{~T}_{i}$, where the statistics $T_{i}$ are identically distributed and also the pairs $T_{i}, T_{j}$ are identically distributed. Thus, we get

$$
\begin{aligned}
\mathrm{E}\left(\sum_{i=1}^{m-r} \mathrm{~T}_{i}\right) & =(m-r) \mathrm{E}\left(\mathrm{~T}_{1}\right) \\
\operatorname{Var}\left(\sum_{i=1}^{m-r} \mathrm{~T}_{i}\right) & =(m-r) \operatorname{Var}\left(\mathrm{T}_{1}\right)+(m-r)(m-r-1) \operatorname{Cov}\left(\mathrm{T}_{1}, \mathrm{~T}_{2}\right)
\end{aligned}
$$

Solve for $\mathrm{E}\left(\mathrm{T}_{1}\right)$ and $\operatorname{Var}\left(\mathrm{T}_{1}\right)$ when $m=p$ and insert above to get

$$
\begin{align*}
\mathrm{E}\left(\sum_{i=1}^{m-r} \mathrm{~T}_{i}\right) & =\frac{m-r}{p-r} \mathrm{E}\left(\sum_{i=1}^{p-r} \mathrm{~T}_{i}\right)  \tag{29}\\
\operatorname{Var}\left(\sum_{i=1}^{m-r} \mathrm{~T}_{i}\right) & =\frac{m-r}{p-r} \operatorname{Var}\left(\sum_{i=1}^{p-r} \mathrm{~T}_{i}\right)-(m-r)(p-m) \operatorname{Cov}\left(\mathrm{T}_{1}, \mathrm{~T}_{2}\right) . \tag{30}
\end{align*}
$$

Thus, it suffices to approximate the moments of the full sample distributions through simulation. Numerically, it appears that better approximations arise when approximating shape and scale parameters instead of mean and variance. We therefore write

$$
\begin{equation*}
\operatorname{Var}\left(\sum_{i=1}^{p-r} \mathrm{~T}_{i}\right)=\delta_{p-r}^{2} \lambda_{p-r}, \quad \mathrm{E}\left(\sum_{i=1}^{p-r} \mathrm{~T}_{i}\right)=\delta_{p-r} \lambda_{p-r} \tag{31}
\end{equation*}
$$

From this, we get the shape and scale parameters as

$$
\frac{1}{\lambda_{p-r}}=\frac{\operatorname{Var}\left(\sum_{i=1}^{p-r} \mathrm{~T}_{i}\right)}{\left\{\mathrm{E}\left(\sum_{i=1}^{p-r} \mathrm{~T}_{i}\right)\right\}^{2}}, \quad \delta_{p-r}=\frac{\operatorname{Var}\left(\sum_{i=1}^{p-r} \mathrm{~T}_{i}\right)}{\mathrm{E}\left(\sum_{i=1}^{p-r} \mathrm{~T}_{i}\right)} .
$$

Hence, we simulated $\lambda_{p-r}, \delta_{p-r}$ and $\operatorname{Cov}\left(\mathrm{T}_{1}, \mathrm{~T}_{2}\right)$ and constructed response surfaces to approximate the distribution of $\mathrm{DF}_{\ell}(m-r, p-r ; v)$. Following JMN and Doornik (1998), we applied a variety of data generating processes and present the results using response surface analysis.

The quantities $\lambda_{p-r}, \delta_{p-r}$ and $\operatorname{Cov}\left(\mathrm{T}_{1}, \mathrm{~T}_{2}\right)$ were simulated for a set of given $p-r, T$ and relative break points. Following JMN, we chose $q=3$ as the maximum number of sub-samples, with $a$ and $b$ representing the smallest and the second smallest of relative sample lengths, respectively. For example, if $q=2$ along with $v_{1} \leq 1-v_{1}$, we then have $a=0$ and $b=v_{1}$. The grid points $a$ and $b$ were selected in the same way as those for Figure 1 in JMN e.g., $(a, b)=(0,0),(0,0.05),(0,0.1), \cdots$, so that they were subject to the constraints of $a \leq b$ and $b \leq(1-a-b)$ and the total number of their combinations was 20 , along with the selection of non-stationary components $p-r=1, \ldots, 8$. For the overall sample sizes or $T \mathrm{~s}$, JMN used 10 integers derived from $500 / i$ for $i=1, \ldots, 10$, but we quadrupled them in order to improve approximations to the underlying limit distributions of the response variables. Thus, we obtained a new set of 10 sample sizes, $T \mathrm{~s}$, ranging from 200 to 2000. For $\log \lambda_{p-r}$ and $\log \delta_{p-r}$, this simulation design led to $1600(=20 \times 8 \times 10)$ cases, while the number of cases was reduced to 1400 for $\operatorname{Cov}\left(\mathrm{T}_{1}, \mathrm{~T}_{2}\right)$ as a result of missing values corresponding to $p-r=1$.

The computational algorithm used in our study was based on Theorem 4. These asymptotic results justify simulating three sets of $T$-step random walks for broken linear-trend and constant cases and scaling them according to the pre-specified relative sample lengths. The number of simulation replications $N$ was set at 100,000.

For the response surface analysis, we used $\log \lambda_{p-r}, \log \delta_{p-r}$ and $\operatorname{Cov}\left(\mathrm{T}_{1}, \mathrm{~T}_{2}\right)$ as the response variables, instead of the logged means and variance as in JMN. It turns out that the use of these response variables ( $\log \lambda_{p-r}$, in particular) mitigates the residual heteroscedasticity problem, hence resulting in a reduction of the number of indicator variables required for $p-r=2$ and $p-r=1$. Note that $\operatorname{Cov}\left(T_{1}, T_{2}\right)$ needs to be included in the set of response variables in any response surface study, in order to make use of Equation (30). In addition, note that taking the $\log$ of $\operatorname{Cov}\left(T_{1}, T_{2}\right)$ is not permissible, since covariance is not always positive.

Compared to JMN, we increased the maximum number of observations from $T=500$ to $T=2000$. It was found that the large-sample $(T \geq 1000)$ approximates of the mean and variance in small dimensions ( $p-r \leq 3$ ) tend to be rather different from those when $T$ is small. This finding is consistent with Doornik (1998), who introduced a set of indicator variables being assigned 1 for $p-r=2$ and $p-r=1$ and assigned 0 otherwise; these indicators put residual heteroscedasticity under control even in the presence of influential values for $p-r=2$ and $p-r=1$.

We regressed each of the three response variables, $\log \lambda_{p-r}, \log \delta_{p-r}$ and $\operatorname{Cov}\left(\mathrm{T}_{1}, \mathrm{~T}_{2}\right)$ on a set of regressors formed from $a, b, p-r$ and $T$. Our baseline function form was a modified version of Equation (3.11) in JMN. In the context of the present paper, the equation in JMN is expressed as

$$
y=\sum_{m=0}^{2}\left(\phi_{m}+\sum_{i=1}^{4} \varphi_{i m} z_{i}+\sum_{i=1}^{4} \sum_{j \geq i} \psi_{i j m} z_{i} z_{j}+\sum_{i=1}^{4} \sum_{j \geq i} \sum_{k \geq j} \omega_{i j k m} z_{i} z_{j} z_{k}\right) d_{m}
$$

where $y$ is either $\log \lambda_{p-r}, \log \delta_{p-r}$ or $\operatorname{Cov}\left(\mathrm{T}_{1}, \mathrm{~T}_{2}\right)$, while $z_{1}=p-r, z_{2}=a, z_{3}=b, z_{4}=T^{-1}$ and $d_{m}=(p-r)^{-m}$. Following Doornik (1998), we also added to this equation a set of indicator variables as explanatory variables, each of which is 1 for a selected value of dimension $p-r$ and is 0 otherwise.

Performing a series of regression analyses and carefully removing insignificant explanatory variables by utilising the Autometrics option available in PcGive (Doornik and Hendry 2013), we arrived at parsimonious response-surface functions for $\log \lambda_{p-r}, \log \delta_{p-r}$ and $\operatorname{Cov}\left(\mathrm{T}_{1}, \mathrm{~T}_{2}\right)$; these functions are henceforth denoted $f_{p-r}^{z}(p-r, a, b, T)$ with $z$ taking values $\lambda, \delta$ and cov, respectively.

Tables A1 and A2 in Appendix A record the rounded coefficients for $a, b, p-r$ and their variants in the response surface regression for the broken linear trend case and the broken constant case, respectively. The inverse of the observation number, $T^{-1}$, and its variants such as $T^{-2}$, also play critical roles in the response surface regression, but all of them are irrelevant asymptotically and thus disregarded when calculating the limit approximates based on these tables.

It should also be noted that a response-surface regression analysis of $\operatorname{Cov}\left(T_{1}, T_{2}\right)$ was technically difficult in terms of residual diagnostic tests. Doornik (1998) used the average of estimates for $\operatorname{Cov}\left(T_{1}, T_{2}\right)$ when performing a response surface analysis for partial systems with no break. We adhered to the regression approach, rather than simply taking the average of the covariance estimates, by assigning importance to various significant influences of $a, b$ and $p-r$ on the behaviour of $\operatorname{Cov}\left(T_{1}, T_{2}\right)$. This regression analysis indeed bore fruit and clarified the highly complex structure of the dependence of $\operatorname{Cov}\left(\mathrm{T}_{1}, \mathrm{~T}_{2}\right)$ on $a, b, p-r$ and its variants, as shown in the third column of each of Tables A1 and A2. These findings about $\operatorname{Cov}\left(T_{1}, T_{2}\right)$ are not known in the literature, thus giving added value to the response surface study conducted in this paper, although the impact of variation in $\operatorname{Cov}\left(T_{1}, T_{2}\right)$ on the approximate shape and scale parameters may not always be large.

Tables 1 and 2 display a set of examples demonstrating the accuracy of the response surface regression results. A class of approximately $95 \%$ limit quantiles is presented in each of the tables for various combinations of $a, b, p-r$ and $m-r$, when either broken-linear-trend or broken-constant specifications are adopted in analysis. Approximate quantiles in the fifth column $\left(q_{95}\right)$ in Tables 1 and 2 are derived from Tables A1 and A2, respectively; that is, they are from the full-system-based response surface analysis, combined with the mappings (29), (30) and (31). By contrast, approximate quantiles recorded in the sixth column $\left(q_{95}^{*}\right)$ of each table, except those for $a=b=0$, were obtained directly from auxiliary response surface regressions based on partial-system simulations with the same Ts and $N$ as above. Each of these auxiliary regression equations employed a simulated $95 \%$ quantile as a response variable and involved a constant, $T^{-1}$ and its powers if necessary, as explanatory variables. The regression equations vary in specification for the purpose of capturing the underlying smooth response surfaces of various simulated quantiles; the graph of each regression's actual and fitted values was checked to ensure the capturing of the underlying smoothness. Estimated constants in these regression equations are recorded in the columns for $q_{95}^{*}$ as approximate $95 \%$ limit quantiles. The limit quantiles in $q_{95}^{*}$ for $a=b=0$ (that is, no break cases) were taken from Doornik (2003).

Tables 1 and 2 show that the quantiles in $q_{95}$ almost coincide with those in $q_{95}^{*}$ regardless of specifications of the deterministic terms; see the seventh column of each table for $\left|q_{95} / q_{95}^{*}-1\right|$, a series of absolute relative errors, all of which are very small. This correspondence can be seen as strong evidence supporting the validity of the proposed approximation method based on the full model. Furthermore, the eighth column of each table records a class of discrepancies in approximate $p$-values, defined as $\Delta \mathrm{p}^{a p p}=g\left(q^{95}\right)\left(q^{95}-q_{95}^{*}\right)$, in which $g(\cdot)$ represents a gamma density function calculated from simulated mean and variance. Most of the discrepancies are very small, and even the largest one is around 0.02 when $p-r$ is relatively large, for which we should recall that a large value of $p-r$ could give rise to various other distortion issues in practice. The overall evidence allows us to argue that the approximate quantiles work as useful critical values in applications from a practical viewpoint. The Supplementary Materials includes an Ox code for simulating asymptic distribution. This can be used if further precision is needed.

As a caveat in relation to large values for $p-r$, let us recall that our response surface regression was conducted by using a class of realistic number of non-stationary variables, $p-r=1, \ldots, 8$, which suffice in most applied research. Thus, an empirical study using a partial system of large
dimension may require careful examinations of the underlying cointegrating rank, in addition to the application of the proposed PLR tests to the data under study, as discussed by Juselius $(2006, \S 8)$.

Table 1. A comparative analysis of $95 \%$ limit quantiles: broken linear-trend models.

| $p-r$ | $m-r$ | $a$ | $b$ | $q_{95}$ | $q_{95}^{*}$ | $\left\|q_{95} / q_{95}^{*}-\mathbf{1}\right\|$ | $\Delta \mathbf{p}^{a p p}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | ---: |
| 2 | 1 | 0.0 | 0.0 | 15.45 | 15.33 | 0.0078 | 0.0058 |
| 2 | 1 | 0.0 | 0.3 | 21.25 | 21.25 | 0.0000 | 0.0000 |
| 2 | 1 | 0.1 | 0.4 | 25.63 | 25.76 | 0.0050 | -0.0065 |
| 2 | 1 | 0.2 | 0.3 | 27.23 | 27.11 | 0.0044 | 0.0055 |
| 2 | 1 | 0.3 | 0.3 | 27.74 | 27.62 | 0.0043 | 0.0056 |
| 4 | 3 | 0.0 | 0.0 | 50.29 | 50.08 | 0.0042 | 0.0100 |
| 4 | 3 | 0.0 | 0.3 | 65.09 | 64.97 | 0.0018 | 0.0057 |
| 4 | 3 | 0.1 | 0.4 | 77.01 | 76.84 | 0.0022 | 0.0082 |
| 4 | 3 | 0.2 | 0.3 | 80.25 | 80.11 | 0.0017 | 0.0068 |
| 4 | 3 | 0.3 | 0.3 | 81.92 | 81.84 | 0.0010 | 0.0039 |
| 5 | 3 | 0.0 | 0.0 | 57.35 | 57.32 | 0.0005 | 0.0015 |
| 5 | 3 | 0.0 | 0.3 | 72.27 | 72.03 | 0.0033 | 0.0112 |
| 5 | 3 | 0.1 | 0.4 | 84.00 | 83.98 | 0.0002 | 0.0010 |
| 5 | 3 | 0.2 | 0.3 | 87.23 | 87.10 | 0.0015 | 0.0063 |
| 5 | 3 | 0.3 | 0.3 | 88.44 | 88.47 | 0.0003 | -0.0015 |
| 7 | 4 | 0.0 | 0.0 | 91.64 | 91.79 | 0.0016 | -0.0077 |
| 7 | 4 | 0.0 | 0.3 | 110.97 | 110.81 | 0.0014 | 0.0076 |
| 7 | 4 | 0.1 | 0.4 | 126.33 | 126.34 | 0.0001 | -0.0005 |
| 7 | 4 | 0.2 | 0.3 | 130.53 | 130.07 | 0.0035 | 0.0215 |
| 7 | 4 | 0.3 | 0.3 | 131.26 | 131.45 | 0.0014 | -0.0097 |

Notes. $q_{95}$ denotes $95 \%$ limit quantiles approximated from the full systems, while $q_{95}^{*}$ denotes those calculated directly from the partial systems. $\Delta \mathrm{p}^{a p p}$ represents discrepancies in approximate $p$-values.

Table 2. A comparative analysis of $95 \%$ limit quantiles: broken constant models.

| $p-r$ | $m-r$ | $a$ | $b$ | $q_{95}$ | $q_{95}^{*}$ | $\left\|q_{95} / q_{95}^{*}-\mathbf{1}\right\|$ | $\Delta \mathbf{p}^{a p p}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | ---: |
| 2 | 1 | 0.0 | 0.0 | 12.21 | 12.28 | 0.0057 | -0.0036 |
| 2 | 1 | 0.0 | 0.3 | 15.51 | 15.55 | 0.0026 | -0.0020 |
| 2 | 1 | 0.1 | 0.4 | 18.24 | 18.35 | 0.0060 | -0.0055 |
| 2 | 1 | 0.2 | 0.3 | 18.71 | 18.75 | 0.0021 | -0.0020 |
| 2 | 1 | 0.3 | 0.3 | 18.81 | 18.87 | 0.0032 | -0.0030 |
| 4 | 3 | 0.0 | 0.0 | 42.76 | 42.60 | 0.0038 | 0.0077 |
| 4 | 3 | 0.0 | 0.3 | 50.66 | 50.71 | 0.0010 | -0.0024 |
| 4 | 3 | 0.1 | 0.4 | 57.40 | 57.67 | 0.0047 | -0.0141 |
| 4 | 3 | 0.2 | 0.3 | 58.63 | 58.69 | 0.0010 | -0.0029 |
| 4 | 3 | 0.3 | 0.3 | 58.83 | 58.90 | 0.0012 | -0.0035 |
| 5 | 3 | 0.0 | 0.0 | 50.06 | 49.96 | 0.0020 | 0.0049 |
| 5 | 3 | 0.0 | 0.3 | 57.88 | 57.73 | 0.0026 | 0.0073 |
| 5 | 3 | 0.1 | 0.4 | 64.64 | 64.73 | 0.0014 | -0.0046 |
| 5 | 3 | 0.2 | 0.3 | 65.66 | 65.50 | 0.0024 | 0.0078 |
| 5 | 3 | 0.3 | 0.3 | 65.62 | 65.66 | 0.0006 | -0.0020 |
| 7 | 4 | 0.0 | 0.0 | 82.47 | 82.35 | 0.0015 | 0.0059 |
| 7 | 4 | 0.0 | 0.3 | 92.22 | 92.26 | 0.0004 | -0.0020 |
| 7 | 4 | 0.1 | 0.4 | 101.46 | 101.56 | 0.0010 | -0.0050 |
| 7 | 4 | 0.2 | 0.3 | 102.01 | 102.04 | 0.0003 | -0.0015 |
| 7 | 4 | 0.3 | 0.3 | 101.81 | 102.16 | 0.0034 | -0.0182 |

Notes. $q_{95}$ denotes $95 \%$ limit quantiles approximated from the full systems, while $q_{95}^{*}$ denotes those calculated directly from the partial systems. $\Delta \mathrm{p}^{a p p}$ represents discrepancies in approximate $p$-values.

### 4.2. Implementation of Response Surface

The response surface in Tables A1 and A2 are used as follows. The response surface is aimed at the situation with two breaks. However, Theorem 4 shows that with a simple correction the response surface can also be used with a single break or no break.

In the case of $q=3$ sample periods and thus 2 breaks at $T_{1}, T_{2}$, where $0<T_{1}<T_{2}<T$, we let $a, b$ be the smallest and second-smallest relative sub-sample length. Thus, if $v_{1}=T_{1} / T, v_{2}=\left(T_{2}-T_{1}\right) / T$, $v_{3}=\left(T-T_{2}\right) / T$ so that $v_{1}+v_{2}+v_{3}=1$. We choose $a=\min \left(v_{1}, v_{2}, v_{3}\right)$ and $b=\operatorname{median}\left(v_{1}, v_{2}, v_{3}\right)$.

In the case of $q=2$ sample periods and thus 1 break at $T_{1}$, where $0<T_{1}<$, then $v_{1}=T_{1} / T$, $v_{2}=\left(T-T_{1}\right) / T$, so that $v_{1}+v_{2}=1$. We let $a=0$ and $b=\min \left(v_{1}, v_{2}\right)$.

In the case of $q=1$ sample period and thus no break, let $a=b=0$. Theorem 4 and (28) show that the mean and variance for the cases where $q<3$ can be found from those for $q=3$ by choosing $a, b$ as indicated and subtracting $(3-q)(p-r)$ and $2(3-q)(p-r)$, respectively.

Given the choices of $p-r, m-r, a$ and $b$, compute the approximations to

$$
\begin{equation*}
f_{p-r}^{\lambda}=\log \lambda_{p-r}, \quad f_{p-r}^{\delta}=\log \delta_{p-r}, \quad c_{p-r}=\operatorname{Cov}\left(\mathrm{T}_{1}, \mathrm{~T}_{2}\right) \tag{32}
\end{equation*}
$$

Table A1 is used for the case with a broken linear trend while Table A2 is used for the case with a broken constant. This is then inserted in (31), which in turn is inserted into (29), (30), while correcting for the number of breaks, that is,

$$
\begin{align*}
\mathrm{E}\left(\sum_{i=1}^{m-r} \mathrm{~T}_{i}\right) & =\frac{m-r}{p-r} \exp \left(f_{p-r}^{\delta}+f_{p-r}^{\lambda}\right)-(3-q)(m-r)  \tag{33}\\
\operatorname{Var}\left(\sum_{i=1}^{m-r} \mathrm{~T}_{i}\right) & =\frac{m-r}{p-r} \exp \left(2 f_{p-r}^{\delta}+f_{p-r}^{\lambda}\right)-(m-r)(p-m) c_{p-r}-2(3-q)(m-r) \tag{34}
\end{align*}
$$

Finally, we approximate the quantile of interest or the $p$-value of the observed $P L R$ statistic using a gamma distribution with mean and variance matching (33) and (34). Equivalently, one can specify the shape and scale of the gamma distribution as mean ${ }^{2} /$ variance and variance/mean.

A spreadsheet for implementing the response surfaces in Tables A1 and A2 is available in the Supplementary Materials. This also includes an Ox program for simulating the asymptotic distributions and calculating $p$-values of observed test statistics for specifications outside the range covered by Tables A1 and A2, for instance when the number of structural breaks is greater than 2 or $q>3$.

## 5. Empirical Illustration

As empirical illustration, we analyse a set of quarterly time series data from Schreiber (2015), who attained an econometric system for the exchange rate and bilateral trade between the UK and Germany. She decomposed the UK-Germany economic system into two blocks, a foreign exchange block and bilateral trade block, in order to obtain a data-congruent representation useful for forecasting and policy analysis. Various econometric studies were conducted by Schreiber (2015), and one of them was the analysis of a partial model for the bilateral trade block with a structural break. The methodology developed in the above sections enables us to conduct formal tests for cointegrating rank that underlies such a partial system subject to a break. This partial system analysis may also be encouraged in terms of local power advantage of partial-system-based tests over those based on a full system under weak exogeneity, as demonstrated by Doornik et al. (1998) as well as Kurita (2011).

Figure 1 presents an overview of the quarterly data spanning the sample period of the first quarter in 1991-the second quarter in 2014, denoted as 1991.1-2014.2 hereafter. The variable $t b_{t}$ is the trade balance between the UK and Germany, i.e., the difference between the log of exports of UK goods to Germany and the log of imports of German goods to the UK; dulc $c_{t}$ represents the unit labour cost differential between the two countries; $y_{t}$ and $y_{t}^{*}$ denote the logs of the UK and German gross domestic products, respectively; $p p p_{t}$ represents the terms of trade in logarithm. See Schreiber (2015) for further details of the data. The figure indicates the presence of a structural break around 2008-2009 attributable to a global economic recession over this period.


Figure 1. Data. (a) $t b_{t}$ is the trade balance between the UK and Germany; (b) $d u l c_{t}$ is the unit labour cost differential between the UK and Germany; (c) $y_{t}$ and $y_{t}^{*}$ are the logs of the UK and German gross domestic products, respectively; (d) $p p p_{t}$ is the terms of trade.

In this empirical illustration, we analyse the data using a bivariate partial autoregressive model for $t b_{t}$ and $d u c_{t}$, with $y_{t}, y_{t}^{*}$ and $p p p_{t}$ assumed to be weakly exogenous for the class of parameters of interest such as cointegrating vectors; that is, $p=5$ and $m=2$. This assumption is based on Schreiber's study, suggesting that modelling the bilateral trade block centering on $t b_{t}$ and $d u l c_{t}$ appears to be conformable to the underlying data structure. The lag-length $k=2$ is selected for our bivariate partial autoregressive model.

With regard to the issue on a structural break, we adopt a broken trend specification; that is, the presence of a shift in the restricted trend as well as the unrestricted constant. The second sub-sample period starts in 2008.3, corresponding to the observation point in $T_{q-1}$ for $q=2$, which results in the selection of relative break points $a=0$ and $b=0.255$. According to (10), our bivariate model with $k=2$ requires a set of two impulse dummy variables for the initial values of the second sub-sample periods. In addition, a pair of impulse dummy variables, $D p 1998(1)$ and $D p 2006(2)$, is employed in our model to capture outliers in the data, as in Schreiber (2015); the former variable is 1 in 1998.1 and zero otherwise for an outlier due to the Asian financial crisis, while the latter is 1 in 2006.2 and zero otherwise, corresponding to an outlier attributable to an increase in oil prices.

A set of residual diagnostic tests for the partial system is reported in Table 3. Most of the test statistics are given in the form $\mathrm{F}_{j}(d f 1, d f 2)$, which denotes an approximate $F$ test (with relevant degrees of freedom $d f 1$ and $d f 2$ ) against the alternative hypothesis $j$. The alternative hypotheses are specified as: 5th-order serial correlation ( $\mathrm{F}_{\text {AR5 }}$, Godfrey 1978), 4th-order autoregressive conditional heteroscedasticity ( $\mathrm{F}_{\text {ARCH4 }}$, Engle 1982), heteroscedasticity ( $\mathrm{F}_{H E T}$, White 1980). Chi-squared tests for normality ( $\chi_{N D}^{2}$, Doornik and Hansen 2008) are also recorded in the table. We also note the following caveats based on recent advances in the field of mis-specification tests: Nielsen (2006) demonstrated that $\mathrm{F}_{\text {AR5 }}$ is a valid test in the presence of unit roots; Berenguer-Rico and Wilms (2018) showed that $\mathrm{F}_{H E T}$ is valid after eliminating outliers from the observations, while $\chi_{N D}^{2}$ is not necessarily valid after the removal of outliers, which was demonstrated by Berenguer-Rico and Nielsen (2017). In any case, no evidence is found in Table 3, suggesting significant mis-specification problems. We can thus judge this partial system is formulated sufficiently well to be subjected to PLR tests for cointegrating rank.

Table 3. Diagnostic test statistics for the estimated partial system.

| Single-Eq. Tests | $\boldsymbol{t} \boldsymbol{b}_{\boldsymbol{t}}$ | dulc $_{\boldsymbol{t}}$ | Vector Tests |  |
| :---: | :---: | :---: | :---: | :---: |
| $\mathrm{F}_{\text {AR5 }}(5,66)$ | $0.946[0.457]$ | $0.777[0.570]$ | $\mathrm{F}_{A R 5}(20,120)$ | $0.542[0.943]$ |
| $\mathrm{F}_{A R C H 4}(4,84)$ | $0.469[0.758]$ | $0.511[0.728]$ | $\mathrm{F}_{H E T}(93,162)$ | $0.838[0.825]$ |
| $\mathrm{F}_{H E T}(31,56)$ | $0.726[0.831]$ | $1.016[0.468]$ | $\chi_{N D}^{2}(4)$ | $2.233[0.693]$ |
| $\chi_{N D}^{2}(2)$ | $1.341[0.512]$ | $0.426[0.808]$ |  |  |
| Notes. Figures in square brackets are $p$-values. |  |  |  |  |

Table 4 presents a class of $P L R$ test statistics for the determination of cointegrating rank, along with the corresponding $p$-values and approximate $95 \%$ limit quantiles calculated from the response surface outcomes in the previous section. We used Table A1 in Appendix A to calculate approximates to $\log \delta_{p-r}, \log \lambda_{p-r}$ and $\operatorname{Cov}\left(\mathrm{T}_{1}, \mathrm{~T}_{2}\right)$, and then applied them to the mappings (29) and (30) adjusted for extra $\chi^{2}$ terms, so that the gamma-distribution approximation method yielded the $p$-values. Table 4 shows that, at the $5 \%$ level, the null hypothesis $r=0$ is rejected while the hypothesis $r \leq 1$ fails to be rejected. Hence, this formal analysis enables us to reach the conclusion of $r=1$, which supports the informal analysis of Schreiber (2015).

Table 4. Testing for cointegrating rank.

|  | $r=\mathbf{0}$ | $r \leq \mathbf{1}$ |
| :--- | :--- | :--- |
| $P L R\left\{H_{\ell}(r) \mid H_{\ell}(2)\right\}$ | $56.610[0.014]^{*}$ | $21.964[0.148]$ |
| $95 \%$ limit quantiles | 50.864 | 26.334 |

Notes. Figures in square brackets are $p$-values. * denotes significance at the $5 \%$ level.
The estimated cointegrating relationship under some additional restrictions is

$$
\begin{equation*}
t b=\underset{(0.121)}{0.259 d u l c_{t}}-\underset{(0.3)}{0.726 p p p_{t}}+\underset{(0.323)}{2.34}\left(y_{t}^{*}-y_{t}\right)-\underset{(0.004)}{0.019} 1_{(\geq 2009: 1)}+v_{t} \tag{35}
\end{equation*}
$$

where a figure in brackets under each coefficient is a standard error and $v_{t}$ is a stationary error. The signs of the coefficients in (35) are the same as those in Schreiber (2015)'s cointegrating equation except for $y_{t}^{*}$. The German income $y_{t}^{*}$ was insignificant in her cointegrating relationship and thus removed from it, while, in (35), $y_{t}^{*}$ plays a significant role, along with $y_{t}$. As a result of checking a set of unrestricted estimates for the cointegrating vector, we have arrived at Equation (35), where the coefficients of $y_{t}$ and $y_{t}^{*}$ are restricted to add to zero, while a zero-restriction is placed on the coefficient for $t 1_{(\leq 2008: 2)}$; that is, a linear trend is present only in the second sub-sample period. The PLR test statistic for these restrictions is 3.571 [0.168], in which the figure in square brackets is a $p$-value according to $\chi^{2}(2)$. Thus, the hypothesis of the overall restrictions cannot be rejected at the $5 \%$ level.

There are several interesting aspects of Equation (35) that are worth discussing here. The real income difference between Germany and the UK, $y_{t}^{*}-y_{t}$, has a positive coefficient, implying that a spread in the income difference leads to an improvement in the UK trade balance with Germany. This finding is interpretable in the context of an income effect from each of the two countries. The coefficient for the terms of trade, $p p p_{t}$, should also be noted. It is negative, thus indicating a relative price effect on the trade balance in a theory-consistent manner; that is, a decrease in exports prices relative to import prices leads to trade balance improvement, so that the well-known elasticity approach to trade balance appears to be empirically valid for the two countries. Furthermore, the linear trend $t$ is significant solely in the second sub-sample period, suggesting long-lasting influences of the global recession on the two countries' trade balance and other economic variables.

Finally, we will check that the three variables, $y_{t}, y_{t}^{*}$ and $p p p_{t}$, are indeed weakly exogenous for the class of parameters of interest. We follow the testing procedure suggested by Johansen (1992a), Boswijk (1992) and HJNR. First, the restricted cointegrating combination is added as a regressor to a marginal system (9) for $Z_{t}=\left(y_{t}, y_{t}^{*}, p p p_{t}\right)^{\prime}$. Second, a standard regression analysis is performed to
test for the significance of the cointegrating combination in each equation. Table 5 reports a class of $L R$ test statistics for the exclusion of the empirical cointegrating linkage from each equation in the marginal system. Judging from the reported $p$-values according to $\chi^{2}(1)$, none of the test statistics indicate evidence against the assumption of weak exogeneity; thus, the preceding partial-system analysis of cointegrating rank has been justified.

Table 5. Checking weak exogeneity.

| $y_{t}$ | $y_{t}^{*}$ | $p p p_{t}$ |
| :---: | :---: | :---: |
| $0.004[0.951]$ | $1.183[0.277]$ | $1.954[0.162]$ |

Notes. Figures in square brackets are $p$-values.

## 6. Conclusions

This study has explored partial cointegrated vector autoregressive models subject to structural breaks in deterministic terms, a linear trend and constant. The Granger-Johansen representation of the full model in JMN has been reexamined, leading to a useful clarification of roles of the initial values in asymptotic analysis. A class of log likelihood ratio test statistics for cointegrating rank has then been introduced in the proposed partial-model framework. We have investigated asymptotic theory under a general class of innovation distributions allowing martingale difference sequences with conditional heteroscedasticity. The derived limit distributions of the statistics are closely related to those for the full models investigated by JMN. This relationship allows us to perform a response surface analysis in a simplified full-system framework, instead of relying on laborious partial-system-based simulations. The outcomes of the analysis are summarised as a set of two statistical tables providing valuable information for inference on the underlying cointegrating rank. Lastly, an empirical analysis of real-life data from the UK and Germany has demonstrated the practicality of these tables in applied economic research. As a result of this study, the partial cointegrated models have become more flexible and reliable devices for modelling time series data subject to various structural breaks.

Recently, bootstrap methods have been proposed for cointegration rank testing in full systems (Cavaliere et al. 2012). It would be interesting to extend those to partial systems with or without breaks.

Supplementary Materials: The following are available online at http:/ / www.mdpi.com/2225-1146/7/4/42/s1: A preadsheet for implementing the response surface in Tables A1 and A2, as well as an Ox program for simulating the asymptotic distributions.
Author Contributions: The authors made equal contributions.
Funding: T. Kurita gratefully acknowledges financial support for this work from JSPS KAKENHI (Grant Nos: 15KK0141 and 26380349).
Acknowledgments: We are grateful to the editors and two anonymous referees for constructive comments. The results reported in this paper were presented at Norwegian University of Science and Technology (NTNU), Statistics Norway and the University of Copenhagen. We would like to thank Gunnar Bårdsen, Pål Boug, Søren Johansen and various other seminar participants for helpful comments and suggestions.
Conflicts of Interest: The authors declare no conflict of interest.

## Appendix A. Tables for Response Surfaces

Table A1. Response surfaces for broken trend models.

| $\log \lambda_{p-r}$ | $\log \delta_{p-r}$ |  |  | $\operatorname{Cov}\left(\mathrm{T}_{1}, \mathrm{~T}_{2}\right)$ |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| const. | 4.14 | const. | 0.5987 | const. | -1.298 |
| $(p-r)^{-1}$ | -6.301 | $p-r$ | -0.0538 | $1_{(2)}$ | 0.03616 |
| $(p-r)^{-2}$ | 5.8842 | $a$ | -1.039 | $1_{(4)}$ | -0.027 |
| $(p-r)^{-3}$ | $-2.32576$ | $b$ | -0.39 | $(p-r)^{-3}$ | -2.022 |
| $p-r$ | 0.17 | $(p-r)^{2}$ | 0.00686 | $a$ | -8.689 |
| $a$ | 2.6165 | $a^{2}$ | 5.547 | $b$ | 2.225 |
| $b$ | 2.5245 | $a b$ | 2.331 | $a^{2}$ | 59.77 |
| $(p-r) a$ | -0.0572 | $b^{2}$ | 1.841 | $a b$ | 24.31 |
| $(p-r) b$ | -0.0971 | $(p-r)^{3}$ | -0.00033 | $b^{2}$ | -5.156 |
| $a^{2}$ | -7.550 | $a^{3}$ | -10.42 | $a^{3}$ | -133.5 |
| $a b$ | $-5.323$ | $a b^{2}$ | -4.325 | $a b^{2}$ | -59.05 |
| $b^{2}$ | -7.412 | $b^{3}$ | -2.553 | $a(p-r)^{-1}$ | -29.55 |
| $(p-r)^{3}$ | -0.000124 | $a(p-r)^{-1}$ | 9.905 | $b(p-r)^{-1}$ | -66.58 |
| $(p-r) a b$ | 0.161 | $b(p-r)^{-1}$ | 1.862 | $b^{2}(p-r)^{-1}$ | 255.3 |
| $(p-r) b^{2}$ | 0.179 | $a^{2}(p-r)^{-1}$ | -61.09 | $a^{3}(p-r)^{-1}$ | 280.5 |
| $a^{3}$ | 10.40 | $a b(p-r)^{-1}$ | -17.09 | $a b^{2}(p-r)^{-1}$ | 155.3 |
| $a b^{2}$ | 6.096 | $b^{2}(p-r)^{-1}$ | -11.48 | $b^{3}(p-r)^{-1}$ | -240 |
| $b^{3}$ | 5.851 | $a^{3}(p-r)^{-1}$ | 117.68 | $a(p-r)^{-2}$ | 21.32 |
| $a(p-r)^{-1}$ | -8.860 | $a b^{2}(p-r)^{-1}$ | 35.19 | $b(p-r)^{-2}$ | 71.68 |
| $b(p-r)^{-1}$ | -4.948 | $b^{3}(p-r)^{-1}$ | 18.6 | $b^{2}(p-r)^{-2}$ | -305.7 |
| $a^{2}(p-r)^{-1}$ | 46.15 | $a(p-r)^{-2}$ | -8.836 | $a^{2} b(p-r)^{-2}$ | -321.1 |
| $a b(p-r)^{-1}$ | 31.85 | $b(p-r)^{-2}$ | 1.033 | $b^{3}(p-r)^{-2}$ | 332.1 |
| $b^{2}(p-r)^{-1}$ | 26.12 | $a^{2}(p-r)^{-2}$ | 66.94 | $(p-r) 1_{(3)}$ | 0.038 |
| $a^{3}(p-r)^{-1}$ | -86.58 | $a b(p-r)^{-2}$ | 10.84 | $b^{2} 1_{(3)}$ | -0.184 |
| $a b^{2}(p-r)^{-1}$ | -50.50 | $a^{3}(p-r)^{-2}$ | -140.88 |  |  |
| $b^{3}(p-r)^{-1}$ | -28.78 | $a b^{2}(p-r)^{-2}$ | -30.16 |  |  |
| $a(p-r)^{-2}$ | 5.296 | $b^{3}(p-r)^{-2}$ | -10.05 |  |  |
| $b(p-r)^{-2}$ | 2.386 | $a 1_{(1)}$ | 2.107 |  |  |
| $a^{2}(p-r)^{-2}$ | -29.03 | $b 1_{(1)}$ | -1.029 |  |  |
| $a b(p-r)^{-2}$ | -19.46 | $a^{2} 1_{(1)}$ | -20.63 |  |  |
| $b^{2}(p-r)^{-2}$ | -13.42 | $b^{2} 1_{(1)}$ | 3.511 |  |  |
| $a^{3}(p-r)^{-2}$ | 62.00 | $a^{3} 1_{(1)}$ | 45.85 |  |  |
| $a^{2} b(p-r)^{-2}$ | $-5.880$ | $a b^{2} 1_{(1)}$ | 4.267 |  |  |
| $a b^{2}(p-r)^{-2}$ | 34.59 | $(p-r) b^{2} 1_{(2)}$ | 0.062 |  |  |
| $b^{3}(p-r)^{-2}$ | 15.93 |  |  |  |  |

Note: $1_{(x)}$ is 1 when $p-r=x$ and zero otherwise.

Table A2. Response surfaces for broken constant models.

| $\log \lambda_{p-r}$ | $\log \delta_{p-r}$ |  | $\operatorname{Cov}\left(\mathrm{T}_{1}, \mathrm{~T}_{2}\right)$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| const. | 4.95486 | const. | 0.4472 | const. | -1.531 |
| $(p-r)^{-1}$ | -9.263 | $(p-r)^{-2}$ | 1.17564 | $(p-r)^{-1}$ | 0.9029 |
| $(p-r)^{-2}$ | 9.162 | $(p-r)^{-3}$ | -1.5294 | $a$ | 4.164 |
| $(p-r)^{-3}$ | -3.662 | , | 0.8286 | $(p-r)^{2}$ | 0.01579 |
| $a$ | 3.05 | $(p-r) b$ | -0.0646 | $(p-r) b$ | 0.3388 |
| $b$ | 0.3315 | $a b$ | 1.75 | $a b$ | -27.16 |
| $(p-r)^{2}$ | 0.01738 | $(p-r) b^{2}$ | 0.04051 | $b^{2}$ | -14.15 |
| $(p-r) a$ | -0.128 | $a^{3}$ | -2.084 | $(p-r)^{3}$ | -0.0013 |
| $a^{2}$ | -14.61 | $a b^{2}$ | -3.698 | $(p-r)^{2} b$ | -0.0167 |
| $a b$ | -4.14 | $b^{3}$ | -0.788 | $a^{3}$ | -19.65 |
| $b^{2}$ | -2.419 | $a(p-r)^{-1}$ | -4.819 | $a^{2} b$ | 14.03 |
| $(p-r)^{3}$ | -0.00084 | $b(p-r)^{-1}$ | -3.897 | $a b^{2}$ | 42.2 |
| $(p-r) a^{2}$ | 0.3264 | $a^{2}(p-r)^{-1}$ | 30.49 | $b^{3}$ | 17.43 |
| $(p-r) a b$ | 0.1302 | $a b(p-r)^{-1}$ | -5.108 | $a(p-r)^{-1}$ | -77.72 |
| $(p-r) b^{2}$ | 0.0266 | $b^{2}(p-r)^{-1}$ | 2.273 | $b(p-r)^{-1}$ | -20.52 |
| $a^{3}$ | 21.56 | $a^{3}(p-r)^{-1}$ | -40.9 | $a^{2}(p-r)^{-1}$ | 278.7 |
| $a b^{2}$ | 5.56 | $a b^{2}(p-r)^{-1}$ | 13.37 | $a b(p-r)^{-1}$ | 313.6 |
| $b^{3}$ | 3.03 | $a(p-r)^{-2}$ | 16 | $b^{2}(p-r)^{-1}$ | 169.1 |
| $a(p-r)^{-1}$ | -5.742 | $b(p-r)^{-2}$ | 3.795 | $a^{3}(p-r)^{-1}$ | -461.7 |
| $b(p-r)^{-1}$ | 3.339 | $a^{2}(p-r)^{-2}$ | -110.5 | $a b^{2}(p-r)^{-1}$ | -562.9 |
| $a^{2}(p-r)^{-1}$ | 44.2 | $a^{3}(p-r)^{-2}$ | 184.8 | $b^{3}(p-r)^{-1}$ | -221.2 |
| $a b(p-r)^{-1}$ | 9.66 | $a b^{2}(p-r)^{-2}$ | -4.478 | $a(p-r)^{-2}$ | 81.64 |
| $b^{2}(p-r)^{-1}$ | -4.44 | $(p-r) 1_{(1)}$ | 0.5014 | $a^{2}(p-r)^{-2}$ | -315 |
| $a^{3}(p-r)^{-1}$ | -81.67 | ${ }^{a 1}{ }_{(1)}$ | -9.833 | $a b(p-r)^{-2}$ | -384.8 |
| $a b^{2}(p-r)^{-1}$ | -15.2 | $a^{2} 1_{(1)}$ | 73.02 | $b^{2}(p-r)^{-2}$ | -114.6 |
| $a(p-r)^{-2}$ | 2.41 | $b^{2} 1_{(1)}$ | -5.835 | $a^{3}(p-r)^{-2}$ | 804 |
| $b(p-r)^{-2}$ | -3.44 | $a^{3} 1_{(1)}$ | -130.2 | $a^{2} b(p-r)^{-2}$ | -290 |
| $a^{2}(p-r)^{-2}$ | -24.23 | $b^{3} 1_{(1)}$ | 4.743 | $a b^{2}(p-r)^{-2}$ | 860.7 |
| $b^{2}(p-r)^{-2}$ | 9.6 | $(p-r)^{2} a 1_{(2)}$ | -0.2472 | $b^{3}(p-r)^{-2}$ | 205.2 |
| $a^{3}(p-r)^{-2}$ | 47.34 | $(p-r)^{2} b 1_{(2)}$ | 0.06919 | $b^{2} 1_{(2)}$ | 0.18 |
| $b^{3}(p-r)^{-2}$ | -7.22 | $(p-r) a^{2} 1_{(2)}$ | 3.765 | $(p-r)^{3} 1_{(2)}$ | -0.00017 |
|  |  | $(p-r) b^{2} 1_{(2)}$ | -0.884 | $(p-r) a 1_{(3)}$ | 1.337 |
|  |  | $a^{3} 1_{(2)}$ | -14.06 | $(p-r) b 1_{(3)}$ | -0.0215 |
|  |  | $b^{3} 1_{(2)}$ | 1.944 | $(p-r)^{2} a 1_{(3)}$ | -0.408 |

Note: $1_{(x)}$ is 1 when $p-r=x$ and zero otherwise.

## Appendix B. Proof of the Granger-Johansen Representation

This section provides a proof of Theorem 1, in which the Granger-Johansen representation of the full model with deterministic breaks is presented.

Proof of Theorem 1. The companion form of the homogenous Equation (13) is

$$
\Delta \tilde{X}_{t-1}=\alpha \boldsymbol{\beta}^{\prime} \tilde{X}_{t-1}+\boldsymbol{\iota} \varepsilon_{t}
$$

on the effective sample, see (7). As shown by Hansen (2005, Lemma A.1), $\operatorname{rank}\left(\alpha_{\perp}^{\prime} \Psi \beta_{\perp}\right)=p-r$ stated in Assumption 1 implies that the above homogenous equation is an $I(1)$ system satisfying

$$
\beta^{\prime} \tilde{X}_{t}=\left(I_{r}+\beta^{\prime} \alpha\right) \beta^{\prime} \tilde{X}_{t-1}+\beta^{\prime} \iota \varepsilon_{t} \quad \text { with } \quad\left|\operatorname{eigen}\left(I_{r}+\beta^{\prime} \alpha\right)\right|<1
$$

which is a stable equation (Lai and Wei 1985).

We then follow Kurita and Nielsen (2009) in the analysis of non-stationary components. Start by the homogenous Equation (13) for $1 \leq j \leq q$ and $T_{j-1}+k<t \leq T_{j}$ :

$$
\Delta \tilde{X}_{t}=\alpha \beta^{\prime} \tilde{X}_{t-1}+\sum_{i=1}^{k-1} \Gamma_{i} \Delta \tilde{X}_{t-i}+\varepsilon_{t}
$$

Pre-multiplying the above equation by $\alpha_{\perp}^{\prime}$ and replacing $\Delta \tilde{X}_{t-i}=\Delta \tilde{X}_{t}-\sum_{\ell=0}^{i-1} \Delta^{2} \tilde{X}_{t-\ell}$, we collect repeated terms $\Delta \tilde{X}_{t}$ on the left-hand side to find

$$
\alpha_{\perp}^{\prime} \Psi \Delta \tilde{X}_{t}=-\alpha_{\perp}^{\prime} \sum_{i=1}^{k-1} \Gamma_{i} \sum_{\ell=0}^{i-1} \Delta^{2} \tilde{X}_{t-\ell}+\alpha_{\perp}^{\prime} \varepsilon_{t}
$$

by recalling $\Psi=I_{p}-\sum_{i=1}^{k-1} \Gamma_{i}$. Summing $\alpha_{\perp}^{\prime} \Psi \Delta \tilde{X}_{s}$ over $s=T_{j-1}+k+1, \ldots, t$ yields

$$
\alpha_{\perp}^{\prime} \Psi \tilde{X}_{t}=\alpha_{\perp}^{\prime} \sum_{\substack{s=T_{j-1} \\+k+1}}^{t} \varepsilon_{s}-\alpha_{\perp}^{\prime} \sum_{i=1}^{k-1} \Gamma_{i} \sum_{\ell=0}^{i-1} \Delta \tilde{X}_{t-\ell}-\alpha_{\perp}^{\prime} \Psi \tilde{X}_{T_{j-1}+k}+\alpha_{\perp}^{\prime} \sum_{i=1}^{k-1} \Gamma_{i} \sum_{\ell=0}^{i-1} \Delta \tilde{X}_{T_{j-1}+k-\ell}
$$

Apply the orthogonal projection identity $\alpha_{\perp}^{\prime} \Psi \tilde{X}_{t}=\alpha_{\perp}^{\prime} \Psi \beta_{\perp} \bar{\beta}_{\perp}^{\prime} \tilde{X}_{t}+\alpha_{\perp}^{\prime} \Psi \bar{\beta} \beta^{\prime} \tilde{X}_{t}$ to the left-hand side and then pre-multiply both sides by $\beta_{\perp}\left(\alpha_{\perp}^{\prime} \Psi \beta_{\perp}\right)^{-1}$ to find the $C$ matrix. Shifting $C \Psi \bar{\beta} \beta^{\prime} \tilde{X}_{t}$ to the right hand side, we arrive at

$$
\begin{aligned}
\beta_{\perp} \bar{\beta}_{\perp}^{\prime} \tilde{X}_{t} & =C \sum_{S=T_{j-1}+k+1}^{t} \varepsilon_{s}-C \Psi \bar{\beta} \beta^{\prime} \tilde{X}_{t}-C \sum_{i=1}^{k-1} \Gamma_{i} \sum_{\ell=0}^{i-1} \Delta \tilde{X}_{t-\ell} \\
& -C \Psi \tilde{X}_{T_{j-1}+k}+C \sum_{i=1}^{k-1} \Gamma_{i} \sum_{\ell=0}^{i-1} \Delta \tilde{X}_{T_{j-1}+k-\ell}
\end{aligned}
$$

Adding $\bar{\beta} \beta^{\prime} \tilde{X}_{t}$ on both sides results in

$$
\begin{align*}
\tilde{X}_{t} & =C \sum_{S=T_{j-1}+k+1}^{t} \varepsilon_{S}+(I-C \Psi) \bar{\beta} \beta^{\prime} \tilde{X}_{t}-C \sum_{i=1}^{k-1} \Gamma_{i} \sum_{\ell=0}^{i-1} \Delta \tilde{X}_{t-\ell}  \tag{A1}\\
& -C \Psi \tilde{X}_{T_{j-1}+k}+C \sum_{i=1}^{k-1} \Gamma_{i} \sum_{\ell=0}^{i-1} \Delta \tilde{X}_{T_{j-1}+k-\ell .} .
\end{align*}
$$

Using the notation $\mathrm{Y}_{i}=-\Gamma_{i}-\cdots-\Gamma_{k-1}$ for $1 \leq i \leq k-1$ as well as the matrix $\Lambda$ defined in (14) leads to the first desired result (17).

Next, we move on to the non-homogenous formulation (12), in which $\mu_{j}$ and $\gamma_{j}$ are distinct from zero. Replace $X_{t}$ in (12) with $\tilde{X}_{t}+\tau_{c, j}+\tau_{l, j} t$ and refer to the proof of Theorem 2.1 in JMN to find

$$
\begin{equation*}
\Psi \tau_{l, j}=\alpha \beta^{\prime}\left(\tau_{c, j}-\tau_{l, j}\right)+\mu_{j} \quad \text { and } \quad \beta^{\prime} \tau_{l, j}+\gamma_{j}^{\prime}=0 \tag{A2}
\end{equation*}
$$

Applying (A2) to (12) recovers the homogenous Equation (13), so the above results derived for (13) are all applicable to (12) under (A2). Substituting $\tilde{X}_{t}=X_{t}-\tau_{c, j}-\tau_{l, j} t$ into (A1) yields the desired representation (18).

## Appendix C. Proofs of Asymptotic Results

In this section, we present a high-level assumption which overrides Assumptions 2 and 3 in the subsequent arguments. We then provide some specific lemmas required for proofs of the limit theorems in Section 3. Finally, we proceed to the proofs of Theorems 2 and 3.

We introduce some notation. For a vector $v$, let the outer product be $v^{\otimes 2}=v v^{\prime}$. For a matrix $m$, the spectral norm is $\|m\|^{2}=\max$ eigen $\left(m^{\prime} m\right)$. Note that $\|m\|^{2} \leq \operatorname{tr}\left(m^{\prime} m\right)$.

## Appendix C.1. A High Level Assumption

In order to give proofs of the theorems introduced in this paper, we need a Law of Large Numbers for the approximately stationary components of the full model, while we require a Functional Central

Limit Theorem and a convergence to a stochastic integral for the non-stationary components of the full model. We formulate these as a the following high level assumption and then prove that it is satisfied under Assumptions 1 and 2.

Assumption A1. Let $\varepsilon_{t}$ be a p-dimensional random variables and suppose that Assumption 1 is satisfied. Let $\tilde{X}_{t}$ satisfy the homogenous Equation (13) and define $\boldsymbol{U}_{t-1}$ as

$$
\boldsymbol{U}_{t-1}=\left(\begin{array}{c}
\beta^{\prime} \tilde{X}_{t-1} \\
\Delta \tilde{X}_{t-1} \\
\vdots \\
\Delta \tilde{X}_{t-k+1}
\end{array}\right)
$$

Suppose that

$$
\begin{equation*}
T^{-1 / 2} \max _{1 \leq t \leq T}\left|\boldsymbol{U}_{t}\right|=\mathrm{op}_{\mathrm{P}}(1) \tag{A3}
\end{equation*}
$$

and

$$
T^{-1} \sum_{t=1}^{T}\left(\begin{array}{c}
\varepsilon_{t}  \tag{A4}\\
\boldsymbol{u}_{t-1} \\
1
\end{array}\right)^{\otimes 2} \xrightarrow{\mathrm{P}}\left(\begin{array}{ccc}
\Omega & 0 & 0 \\
0 & \Sigma_{u} & 0 \\
0 & 0 & 1
\end{array}\right)
$$

where $\Omega$ and $\Sigma_{u}$ are positive definite matrices. Furthermore, let $W_{u}$ be a $p$-dimensional Brownian motion with variance $\Omega$. Suppose that, for $0 \leq u \leq 1$,

$$
\begin{equation*}
T^{-1 / 2} \sum_{t=1}^{\operatorname{int}(T u)} \varepsilon_{t} \xrightarrow{\mathrm{D}} W_{u}, \tag{A5}
\end{equation*}
$$

as a process on $(D[0,1])^{p}$ endowed with the Skorokhod metric with common distortion. Finally,

$$
\begin{equation*}
T^{-1} \sum_{t=1}^{T} \sum_{s=1}^{t-1} \varepsilon_{s} \varepsilon_{t}^{\prime} \xrightarrow{\mathrm{D}} \int_{0}^{1} W_{u} d W_{u}^{\prime} . \tag{A6}
\end{equation*}
$$

The next result explores the conditions of Brown (1971). Subsequently, we use this to show that Assumptions 1 and 2 imply Assumption A1.

Lemma A1. Suppose Assumption 2 is satisfied. Then,
(a) $T^{-1} \sum_{t=1}^{T} \varepsilon_{t} \varepsilon_{t}^{\prime} \xrightarrow{P} \Omega$;
(b) $T^{-1} \sum_{t=1}^{T} \mathrm{E}\left\{\varepsilon_{t}^{\prime} \varepsilon_{t} 1_{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>\delta T\right)} \mid \mathcal{F}_{t-1}\right\} \xrightarrow{\mathrm{P}} 0$ for all $\delta>0$;
(c) $T^{-1} \sum_{t=1}^{T} \mathrm{E}\left\{\varepsilon_{t}^{\prime} \varepsilon_{t} 1_{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>\delta T\right)}\right\} \rightarrow 0$ for all $\delta>0$;
(d) $\max _{1 \leq t \leq T}\left|\varepsilon_{t}^{2}\right| / T \xrightarrow{P} 0$.

Proof of Lemma A1. ( $b, c$ ) Brown (1971, Lemma 2) shows that the conditional Lindeberg condition $(b)$ and the marginal Lindeberg condition (c) are equivalent under Assumption 2( $i, i i$ ).

First, suppose Assumption 2(iii.a), so that $\sup _{t \in \mathbb{N}} \mathrm{E}\left\{\varepsilon_{t}^{\prime} \varepsilon_{t} 1_{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>a\right)} \mid \mathcal{F}_{t-1}\right\}=\mathrm{op}_{\mathrm{P}}(1)$ as $a \rightarrow \infty$. Thus, $\forall \xi>0, \exists a_{0}$ and $\forall a \geq a_{0}$, it follows that

$$
\mathrm{P}\left[\sup _{t \in \mathbb{N}} \mathrm{E}\left\{\varepsilon_{t}^{\prime} \varepsilon_{t} 1_{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>a\right)} \mid \mathcal{F}_{t-1}\right\}>\xi\right]<\xi
$$

Thus, given $\delta>0$ and for $\forall T>a_{0} / \delta$, we find

$$
\mathrm{E}\left\{\varepsilon_{t}^{\prime} \varepsilon_{t} 1_{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>\delta T\right)} \mid \mathcal{F}_{t-1}\right\} \leq \mathrm{E}\left\{\varepsilon_{t}^{\prime} \varepsilon_{t} 1_{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>a_{0}\right)} \mid \mathcal{F}_{t-1}\right\}<\xi
$$

so that the conditional Lindeberg condition (b) follows.
Second, suppose Assumption 2(iii.b), so that sup $\operatorname{teN} \mathrm{E}\left|\varepsilon_{t}\right|^{4}<\infty$. Chebychev's inequality gives

$$
\mathrm{E}\left\{1_{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>\delta T\right)}\right\}=\mathrm{P}\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>\delta T\right) \leq \frac{1}{\delta^{2} T^{2}} \mathrm{E}\left|\varepsilon_{t}\right|^{4}
$$

Next, by the Cauchy-Schwarz inequality and Assumption 2(iii.b), we get

$$
\mathrm{E}\left\{\varepsilon_{t}^{\prime} \varepsilon_{t} 1_{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>\delta T\right)}\right\} \leq\left[\mathrm{E}\left|\varepsilon_{t}\right|^{4} \mathrm{E}\left\{1_{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>\delta T\right)}\right\}\right]^{1 / 2} \leq \delta^{-1} T^{-1} \mathrm{E}\left|\varepsilon_{t}\right|^{4} \leq \delta^{-1} T^{-1} \sup _{t \in \mathbb{N}} \mathrm{E}\left|\varepsilon_{t}\right|^{4} \leq \delta^{-1} T^{-1} C
$$

Hence, $T^{-1} \sum_{t=1}^{T} \mathrm{E}\left\{\varepsilon_{t}^{\prime} \varepsilon_{t} 1_{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>\delta T\right)}\right\} \leq \delta^{-1} T^{-1} \mathrm{C} \rightarrow 0$ so that the marginal Lindeberg condition (c) holds.
(a) We define $U_{T}^{2}=T^{-1} \sum_{t=1}^{T} \varepsilon_{t} \varepsilon_{t}^{\prime}$ and $V_{T}^{2}=T^{-1} \sum_{t=1}^{T} \mathrm{E}\left(\varepsilon_{t} \varepsilon_{t}^{\prime} \mid \mathcal{F}_{t-1}\right)$ and show that $\left\|U_{T}^{2}-V_{T}^{2}\right\| \xrightarrow{\mathrm{P}} 0$. Since $U_{T}^{2}-V_{T}^{2}$ is symmetric, the spectral norm equals the spectral radius, thus it suffices to show that $v^{\prime}\left(U_{T}^{2}-V_{T}^{2}\right) v$ vanishes for any linear combination $v$. In turn, it suffices to consider univariate martingale difference sequences $\varepsilon_{t}$.

First, suppose Assumption 2(iii.a) holds, so that $\sup _{t \in \mathbb{N}} \mathrm{E}\left\{\varepsilon_{t}^{2} 1_{\left(\varepsilon_{t}^{2}>a\right)} \mid \mathcal{F}_{t-1}\right\}=\mathrm{op}(1)$ as $a \rightarrow \infty$. We follow an argument inspired by Anderson and Kunitomo (1992, Theorem 2). Hall and Heyde (1980, Theorem 2.23) show that $U_{T}^{2}-V_{T}^{2} \xrightarrow{P} 0$ whenever the Lindeberg condition in part (b) holds and $\sup _{T \in \mathbb{N}} \mathrm{P}\left(V_{T}^{2}>\lambda\right) \rightarrow 0$ as $\lambda \rightarrow \infty$. To prove the tightness condition, note that

$$
V_{T}^{2}=T^{-1} \sum_{t=1}^{T} \mathrm{E}\left\{\varepsilon_{t}^{2} 1_{\left(\varepsilon_{t}^{2} \leq \lambda\right)} \mid \mathcal{F}_{t-1}\right\}+\mathrm{E}\left\{\varepsilon_{t}^{2} 1_{\left(\varepsilon_{t}^{2}>\lambda\right)} \mid \mathcal{F}_{t-1}\right\} \leq \lambda+\sup _{t \in \mathbb{N}} \mathrm{E}\left\{\varepsilon_{t}^{2} 1_{\left(\varepsilon_{t}^{2}>\lambda\right)} \mid \mathcal{F}_{t-1}\right\}
$$

Thus, to analyse the tightness probability bound,

$$
\mathrm{P}\left(V_{T}^{2}>2 \lambda\right) \leq \mathrm{P}\left[\lambda+\sup _{t \in \mathbb{N}} \mathrm{E}\left\{\varepsilon_{t}^{2} 1_{\left(\varepsilon_{t}^{2}>\lambda\right)} \mid \mathcal{F}_{t-1}\right\}>2 \lambda\right]=\mathrm{P}\left[\sup _{t \in \mathbb{N}} \mathrm{E}\left\{\varepsilon_{t}^{2} 1_{\left(\varepsilon_{t}^{2}>\lambda\right)} \mid \mathcal{F}_{t-1}\right\}>\lambda\right] .
$$

This bound is uniform in $T$ so that $\sup _{T \in \mathbb{N}} \mathrm{P}\left(V_{T}^{2}>\lambda\right) \leq \mathrm{P}\left[\sup _{t \in \mathbb{N}} \mathrm{E}\left\{\varepsilon_{t}^{2} 1_{\left(\varepsilon_{t}^{2}>\lambda\right)} \mid \mathcal{F}_{t-1}\right\}>\lambda\right]$, which vanishes by Assumption 2 (iii.a).

Second, suppose Assumption 2(iii.b) holds, so that $\sup _{t \in \mathbb{N}} \mathrm{E} \varepsilon_{t}^{4}<C<\infty$. Let $m_{t}=\varepsilon_{t}^{2}-\mathrm{E}\left(\varepsilon_{t}^{2} \mid \mathcal{F}_{t-1}\right)$. The Chebychev inequality and the uncorrelatedness of martingale differences gives

$$
\mathcal{P}=\mathrm{P}\left(\left|T^{-1} \sum_{t=1}^{T} m_{t}\right|>\epsilon\right) \leq \frac{1}{\epsilon^{2}} \mathrm{E}\left|T^{-1} \sum_{t=1}^{T} m_{t}\right|^{2}=\frac{1}{T^{2} \epsilon^{2}} \mathrm{E} \sum_{t=1}^{T} m_{t}^{2} .
$$

Jensen's inequality shows $\mathrm{E}\left\{\mathrm{E}\left(\varepsilon_{t}^{2} \mid \mathcal{F}_{t-1}\right)\right\}^{2} \leq \mathrm{E}\left\{\mathrm{E}\left(\varepsilon_{t}^{4} \mid \mathcal{F}_{t-1}\right)\right\}=\mathrm{E} \varepsilon_{t}^{4}$. Thus, the inequality $(a+b)^{2} \leq 2\left(a^{2}+b^{2}\right)$ shows that $\mathrm{E} m_{t}^{2} \leq 2\left[\mathrm{E} \varepsilon_{t}^{4}+\mathrm{E}\left\{\mathrm{E}\left(\varepsilon_{t}^{2} \mid \mathcal{F}_{t-1}\right)\right\}^{2}\right] \leq 4 \mathrm{E} \varepsilon_{t}^{4} \leq 4 C<\infty$. Thus, $\mathcal{P} \leq\left(T \epsilon^{2}\right)^{-1} 4 C \rightarrow 0$.
(d) We show $\mathcal{P}_{T}=\mathrm{P}\left(\max _{1 \leq t \leq T}\left|\varepsilon_{t}\right|^{2}>\delta T\right) \rightarrow 0$ for all $\delta>0$. Note that $\mathcal{P}_{T}=\bigcup_{t=1}^{T} \mathrm{P}\left(\left|\varepsilon_{t}\right|^{2}>\delta T\right)$. Boole's inequality gives $\mathcal{P}_{T} \leq \sum_{t=1}^{T} \mathrm{P}\left(\left|\varepsilon_{t}\right|^{2}>\delta T\right)=\sum_{t=1}^{T} \mathrm{E} 1_{\left(\left|\varepsilon_{t}\right|^{2}>\delta T\right)}$. On the set $\left(\left|\varepsilon_{t}\right|^{2}>\delta T\right)$, we get the further bound $\mathcal{P}_{T} \leq \delta^{-1} T^{-1} \sum_{t=1}^{T} \mathrm{E}\left|\varepsilon_{t}\right|^{2} 1_{\left(\left|\varepsilon_{t}\right|^{2}>\delta T\right)}$, which vanishes by part (c).

We then prove that Assumption A1 is satisfied under Assumptions 1 and 2.
Lemma A2. Suppose that $X_{t}$ and $\varepsilon_{t}$ satisfy Assumptions 1 and 2, respectively, while $\tilde{X}_{t}$ solves the homogenous Equation (13). Then, Assumption A1 is satisfied.

Proof of Lemma A2. Note that the process $\boldsymbol{U}_{t}$ equals $\boldsymbol{\beta}^{\prime} \tilde{\boldsymbol{X}}_{t}$, which is studied in Theorem 1. It satisfies Equation (16), which is of the form $\boldsymbol{U}_{t}=\Phi \boldsymbol{U}_{t-1}+F \varepsilon_{t}$ with $\Phi=\left(I_{r}+\boldsymbol{\beta}^{\prime} \boldsymbol{\alpha}\right)$ for $\boldsymbol{r}=r+p(k-1)$ and $F=\beta^{\prime} \iota$, where $\Phi$ has spectral radius less than unity as verified in Theorem 1.

For (A3), we apply Anderson and Kunitomo (1992, Lemma 1). This requires that $\max _{1 \leq t \leq T}\left|\varepsilon_{t}\right|^{2}=$ $\mathrm{o}_{\mathrm{p}}(T)$, which is proved in Lemma A1 using Assumption 2.

For (A4), use Assumption 2(iii.a). We apply Lemma 2 in Anderson and Kunitomo (1992) to show the convergence of the product moment matrix, which requires Assumption 2(ii, iii.a). Assumption 2 states that $\Omega$ is a positive definite matrix, which results in the positive definiteness of $\Sigma_{u}$ by Anderson and Kunitomo (1992, Lemma 3).

For (A4) using Assumption 2(iii.b). We follow Anderson and Kunitomo (1992, Lemma 2) but avoid their truncation argument. We first argue that $\sum_{t=1}^{T} \boldsymbol{U}_{t-1} F \varepsilon_{t}^{\prime}=\mathrm{op}_{\mathrm{P}}(T)$. Since $\boldsymbol{U}_{t-1} F \varepsilon_{t}^{\prime}$ is a martingale difference sequence with second moments due to Assumption 2(ii) and the spectral norm is bounded by the trace, we obtain

$$
\mathcal{E}=\mathrm{E}\left\|\sum_{t=1}^{T} \boldsymbol{U}_{t-1} F \varepsilon_{t}\right\|^{2} \leq \operatorname{Etr} \sum_{s=1}^{T} \sum_{t=1}^{T} \boldsymbol{U}_{t-1} F \varepsilon_{t}^{\prime} \varepsilon_{s} F^{\prime} \boldsymbol{U}_{s-1}^{\prime}=\operatorname{Etr} \sum_{t=1}^{T} \boldsymbol{U}_{t-1} F \varepsilon_{t}^{\prime} \varepsilon_{t} F^{\prime} \boldsymbol{U}_{t-1}^{\prime}
$$

Applying iterated expectations and using that $\max _{1 \leq t \leq T} \mathrm{E}\left|\varepsilon_{t}\right|^{2}$ is bounded by Assumption 2(iii.b) gives $\mathcal{E} \leq C \sum_{t=1}^{T} \mathrm{E}\left|\boldsymbol{U}_{t-1}\right|^{2}$. Noting that $\boldsymbol{U}_{t}=\sum_{j=0}^{t-1} \Phi^{j} F \varepsilon_{t-j}+\Phi^{t} \boldsymbol{U}_{0}$ and using that $\varepsilon_{t}$ is a martingale difference array, we arrive at

$$
\mathrm{E} U_{t-1}^{\prime} \boldsymbol{U}_{t-1}=\mathrm{E} \sum_{j=0}^{t-2} \varepsilon_{t-1-j}^{\prime} F^{\prime}\left(\Phi^{j}\right)^{\prime} \Phi^{j} F \varepsilon_{t-1-j}=\sum_{j=0}^{t-2} \operatorname{tr}\left\{F^{\prime}\left(\Phi^{j}\right)^{\prime} \Phi^{j} F\right\} \mathrm{E}\left|\varepsilon_{t-1-j}\right|^{2}
$$

Using that $\max _{1 \leq t \leq T} \mathrm{E}\left|\varepsilon_{t}\right|^{2}$ is bounded and $\Phi$ has spectral radius less than unity,

$$
\mathrm{E}\left|U_{t-1}\right|^{2} \leq \sum_{j=0}^{\infty} \operatorname{tr}\left\{F^{\prime}\left(\Phi^{j}\right)^{\prime} \Phi^{j} F\right\} \max _{1 \leq t \leq T} \mathrm{E}\left|\varepsilon_{t}\right|^{2} \leq C
$$

As a consequence $\mathcal{E}=\mathrm{O}(T)$ and, by the Markov inequality, $\sum_{t=1}^{T} \boldsymbol{U}_{t-1} F \varepsilon_{t}^{\prime}=\mathrm{op}_{\mathrm{p}}(T)$. Next, we show $T^{-1} \sum_{t=1}^{T} \boldsymbol{U}_{t-1} \boldsymbol{U}_{t-1}^{\prime} \rightarrow \Sigma_{U}$ in probability. Since $\boldsymbol{U}_{t-1}=\Phi \boldsymbol{U}_{t-2}+F \varepsilon_{t-1}$, then

$$
\begin{aligned}
T^{-1} \sum_{t=1}^{T} \boldsymbol{u}_{t-1} \boldsymbol{u}_{t-1}^{\prime}= & \Phi T^{-1} \sum_{t=1}^{T} \boldsymbol{u}_{t-2} \boldsymbol{U}_{t-2}^{\prime} \Phi^{\prime}+T^{-1} \sum_{t=1}^{T} F \varepsilon_{t-1} \varepsilon_{t-1}^{\prime} F^{\prime} \\
& +\Phi T^{-1} \sum_{t=1}^{T} \boldsymbol{U}_{t-2} \varepsilon_{t-1}^{\prime} F^{\prime}+T^{-1} \sum_{t=1}^{T} F \varepsilon_{t-1} \boldsymbol{u}_{t-2}^{\prime} \Phi^{\prime}
\end{aligned}
$$

Here, the second term converges to $F \Omega F^{\prime}$ by Assumption 2(i) while the last two terms vanish. Since $\max _{1 \leq t \leq T}\left|\boldsymbol{U}_{t}\right|=\mathrm{op}_{\mathrm{P}}\left(T^{1 / 2}\right)$, we therefore get

$$
T^{-1} \sum_{t=1}^{T} \boldsymbol{u}_{t-1} \boldsymbol{U}_{t-1}^{\prime}=\Phi T^{-1} \sum_{t=1}^{T} \boldsymbol{u}_{t-1} \boldsymbol{u}_{t-1}^{\prime} \Phi^{\prime}+F \Omega F^{\prime}+\mathrm{op}(1)
$$

This is a linear equation in $T^{-1} \Sigma_{t=1}^{T} \boldsymbol{U}_{t-1} \boldsymbol{U}_{t-1}^{\prime}$ so that $T^{-1} \Sigma_{t=1}^{T} \boldsymbol{U}_{t-1} \boldsymbol{U}_{t-1}^{\prime} \rightarrow \Sigma_{U}$ in probability where $\Sigma_{U}$ solves $\Sigma_{U}=\Phi \Sigma_{U} \Phi^{\prime}+F \Omega F^{\prime}$. Anderson and Kunitomo (1992, Lemma 3) show invertibility of $\Sigma_{U}$.

For (A5), the Functional Central Limit Theorem follows from the univariate result of Brown (1971, Theorem 3), equipped with Cramér-Wold device, see Billingsley (1968, Theorem 7.7). Brown's result applies under Assumption 2(i,ii) and either of the Lindeberg condition established in Lemma A1 $(b, c)$ under Assumption 2. When using Brown's result, it is convenient to define the
univariate variables $S_{t}=\sum_{s=1}^{t} \varepsilon_{s}$ and $s_{t}^{2}=\sum_{s=1}^{t} E \varepsilon_{s}^{2}$. Brown is concerned with the continuously embedded random walk through the points $\left(s_{t}^{2} / s_{T}^{2}\right),\left(S_{t} / s_{T}\right)$, while we are concerned with the right continuous random walk that is constant $\left(S_{t} / T^{1 / 2}\right)$ on the half-open intervals $\left[t / T^{1 / 2},(t+1) / T^{1 / 2}\right)$. The two embeddings reconcile since $s_{t}^{2} / s_{T}^{2}=t \sigma^{2}$ for some constant $\sigma^{2}$ by Assumption 2(i) and since $\max _{1 \leq t \leq T}|\varepsilon| / T^{1 / 2}$ vanishes by Lemma A1(d) under Assumption 2.

For (A6), the convergence to a stochastic integral for the univariate case is based on the results of Jakubowski et al. (1989), which was referred to by Kurtz and Protter (1996), while the convergence to a stochastic integral for the multivariate case is based on the results of Kurtz and Protter (1991). For the univariate case, Kurtz and Protter (1996, Theorem 7.1) show that we need to check that the martingale array $M_{t}^{T}=T^{-1 / 2} \sum_{s=1}^{t} \varepsilon_{s}$ is uniformly tight, as required by Jakubowski et al. (1989) or, equivalently, it has uniformly controlled variations. We use Kurtz and Protter (1991, Theorem 2.2), which applies to the multivariate case; see also Hansen (1992, Theorem 2.1). Choose $\delta=\infty$ so that $M_{t}^{T, \delta}=M_{t}^{T}$ in Kurtz and Protter's notation. For each $\alpha>0, T \geq 1$, choose stopping times $\tau^{T, \alpha}=\infty$ so that $\mathrm{P}\left(\tau^{T, \alpha} \leq \alpha\right)=0 \leq 1 / \alpha$. Then, we obtain a quadratic variation processes

$$
\left[M^{T, \delta}\right]_{t \wedge \tau^{T, \alpha}}=T^{-1} \sum_{s=1}^{t \wedge \tau^{T, \alpha}} \varepsilon_{s} \varepsilon_{s}^{\prime} \leq T^{-1} \sum_{t=1}^{T} \varepsilon_{t} \varepsilon_{t}^{\prime}=\left[M^{T, \delta}\right]_{T}
$$

so that $\mathrm{E}\left[M^{T, \delta}\right]_{t \wedge \tau^{T, \alpha}} \leq \mathrm{E}\left[M^{T, \delta}\right]_{T}$. From Assumption 2(i), it follows that $\mathrm{E}\left[M^{T, \delta}\right]_{T} \rightarrow \Omega$. Consequently, we have $\sup _{T}\left\|\mathrm{E}\left[M^{T, \delta}\right]_{T}\right\|<\infty$. In turn, $\sup _{T}\left\|\mathrm{E}\left[M^{T, \delta}\right]_{t \wedge \tau^{T, \alpha}}\right\|<\infty$ for each $t$, so that $M_{t}^{T}$ has uniformly controlled variations.

Proof of Lemma 1. Assumption 3 has $\mathrm{E}\left(\varepsilon_{t} \varepsilon_{t}^{\prime} \mid \mathcal{F}_{t-1}\right)=\Omega$ a.s., so that $T^{-1} \sum_{t=1}^{T} \mathrm{E}\left(\varepsilon_{t} \varepsilon_{t}^{\prime} \mid \mathcal{F}_{t-1}\right)=\Omega$ a.s. follows and Assumption 2(ii) holds. Taking iterated expectations, we obtain $\mathrm{E}\left(\varepsilon_{t} \varepsilon_{t}^{\prime}\right)=\Omega$, which leads to $T^{-1} \sum_{t=1}^{T} \mathrm{E}\left(\varepsilon_{t} \varepsilon_{t}^{\prime}\right)=\Omega$ and thus Assumption 2(i) is satisfied. Lastly, we show that Assumption 2(iii.a) is implied by Assumption 3(ii). Using Hölder's inequality, we find, for $\eta=\xi / 2>0$,

$$
\mathrm{E}\left\{\varepsilon_{t}^{\prime} \varepsilon_{t} 1_{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>a\right)} \mid \mathcal{F}_{t-1}\right\} \leq\left[\mathrm{E}\left(\left|\varepsilon_{t}\right|^{2+2 \eta} \mid \mathcal{F}_{t-1}\right)\right]^{1 /(1+\eta)}\left[\mathrm{E}\left\{1_{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>a\right)}^{(1+\eta) / \eta} \mid \mathcal{F}_{t-1}\right\}\right]^{\eta /(1+\eta)}
$$

in which we note the equality $1_{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>a\right)}^{(1+\eta) / \eta}=1_{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>a\right)}$. Hence, writing the expectation of the indicator as a probability, and also using Markov's inequality, we arrive at

$$
\mathrm{E}\left\{1_{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>a\right)}^{(1+\eta) / \eta} \mid \mathcal{F}_{t-1}\right\}=\mathrm{P}\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>a \mid \mathcal{F}_{t-1}\right) \leq \frac{1}{a^{1+\eta}} \mathrm{E}\left\{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}\right)^{1+\eta} \mid \mathcal{F}_{t-1}\right\}
$$

In combination, we obtain $\mathrm{E}\left\{\varepsilon_{t}^{\prime} \varepsilon_{t} 1_{\left(\varepsilon_{t}^{\prime} \varepsilon_{t}>a\right)} \mid \mathcal{F}_{t-1}\right\} \leq a^{-\eta} \mathrm{E}\left(\left|\varepsilon_{t}\right|^{2+2 \eta} \mid \mathcal{F}_{t-1}\right)$, which vanishes as $a \rightarrow \infty$ uniformly in $t$, since the $\mathrm{E}\left(\left|\varepsilon_{t}\right|^{2+2 \eta} \mid \mathcal{F}_{t-1}\right)$ is uniformly bounded by assumption.

Remark A1. We give an example of a martingale difference sequence $\left(\varepsilon_{t}, \mathcal{F}_{t}\right)$ satisfying $\sup _{t \in \mathbb{N}} \mathrm{E}\left|\varepsilon_{t}\right|^{4}<$ $\infty$ but violating Assumption 2(iii.a), i.e., $\sup _{t \in \mathbb{N}} \mathrm{E}\left\{\left|\varepsilon_{t}\right|^{2} 1_{\left(\left|\varepsilon_{t}\right|^{2}>a\right)} \mid \mathcal{F}_{t-1}\right\} \rightarrow 0$ in probability, which is from Anderson and Kunitomo (1992). Consider the probability space $\{(0,1], \mathcal{F}, \mathrm{P}\}$, where $\mathcal{F}$ is the Borel field on $(0,1]$ and $P$ is the uniform distribution. Consider also a dyadic sequence with indices $n=1,2, \ldots$ and $k_{n}=1, \ldots, 2^{n}$, so that $t=\sum_{j=1}^{n-1} 2^{j}+k_{n}$, and define

$$
\varepsilon_{t}(\omega)=\varepsilon_{n k_{n}}(\omega)=\left\{\begin{array}{ll}
n & \text { if } 2 k_{n}-1<2^{n+1} \omega \leq 2 k_{n} \\
-n & \text { if } 2 k_{n}-2<2^{n+1} \omega \leq 2 k_{n}-1 \\
0 & \text { otherwise. }
\end{array}\right\} .
$$

We note that $\mathrm{E} \varepsilon_{t}=0$ while $\mathrm{E} \varepsilon_{t}^{4}=n^{4} / 2^{n}$ is uniformly bounded in $n$. The natural filtration of $\varepsilon_{t}$ is then given by the $\sigma$-fields $\mathcal{F}_{0}=\sigma\{[0,1]\}, \mathcal{F}_{1}=\sigma\left\{\left(0, \frac{1}{4}\right],\left(\frac{1}{4}, \frac{1}{2}\right], \mathcal{F}_{0}\right\}, \mathcal{F}_{2}=\sigma\left\{\left(\frac{1}{2}, \frac{3}{4}\right],\left(\frac{3}{4}, 1\right], \mathcal{F}_{1}\right\}, \mathcal{F}_{3}=$
$\sigma\left\{\left(0, \frac{1}{8}\right],\left(\frac{1}{8}, \frac{1}{4}\right], \mathcal{F}_{2}\right\}$ and so on. We find that $\mathrm{E}\left(\varepsilon_{t} \mid \mathcal{F}_{t-1}\right)=0$ while $\left[\sup _{t \in \mathbb{N}} \mathrm{E}\left\{\left|\varepsilon_{t}\right|^{2} 1_{\left(\left|\varepsilon_{t}\right|^{2}>a\right)} \mid \mathcal{F}_{t-1}\right\}\right](\omega)=\infty$ for all $a$ and all $\omega$, so that the random variables $\sup _{t \in \mathbb{N}} \mathrm{E}\left\{\left|\varepsilon_{t}\right|^{2} 1_{\left(\left|\varepsilon_{t}\right|^{2}>a\right)} \mid \mathcal{F}_{t-1}\right\}$ cannot vanish in probability.

## Appendix C.2. Several Lemmas for the Partial Systems

The asymptotic properties of the product moment matrices, $S_{i j}$ for $i, j=0,1$ defined in (24), are investigated so as to adapt Lemmas 10.1 and 10.3 of Johansen (1995) to the present model. We do this by combining various ideas and techniques from HJNR, JMN and Kurita and Nielsen (2009). These papers assume normal innovations, which we have generalised as Assumptions 2 and 3 in our study. This means that we have to be careful when defining the limits of product moments of various non-integrated components. This issue is addressed in the following lemma:

Lemma A3. Suppose that Assumptions 1 and A1 are satisfied. Let

$$
V_{t}=\left(\begin{array}{c}
\varepsilon_{t} \\
\Delta X_{t} \\
\beta^{\prime} X_{t-1}+\gamma_{j} t
\end{array}\right), \quad Q_{t}=\left(\begin{array}{c}
\Delta X_{t-1} \\
\vdots \\
\Delta X_{t-k+1}
\end{array}\right) \quad \text { and } \quad \bar{Q}_{t}=\binom{Q_{t}}{1}
$$

Let $v_{j}=T_{j} / T$ be relative break points for $j=0, \ldots, q$ and define the sample product moment matrix of $V_{t}$ corrected for $Q_{t}$ and a constant as

$$
M_{V V \cdot Q, 1}=\sum_{j=1}^{q} \frac{1}{T} \sum_{t=T_{j-1}+k}^{T_{j}}\left\{V_{t}-\sum_{s=T_{j-1}+k}^{T_{j}} V_{t} \bar{Q}_{s}^{\prime}\left(\sum_{s=T_{j-1}+k}^{T_{j}} \bar{Q}_{s}^{\otimes 2}\right)^{-1} \bar{Q}_{t}\right\}^{\otimes 2} .
$$

Then, as $T \rightarrow \infty$ with fixed relative break points $v_{j}$, we get that $M_{V V \cdot Q, 1}$ converges in probability to a positive definite matrix with the structure

$$
\left(\begin{array}{ccc}
\Omega & \Omega & 0  \tag{A7}\\
\Omega & \Sigma_{x x} & \Sigma_{x \beta} \\
0 & \Sigma_{\beta x} & \Sigma_{\beta \beta}
\end{array}\right)
$$

where $\Sigma_{x \beta}=\alpha \Sigma_{\beta \beta}$ and $\Sigma_{x x}=\alpha \Sigma_{\beta x}+\Omega$ hold.
Proof of Lemma A3. We start with the homogenous Equation (16). For $1 \leq j \leq q$ and $T_{j-1}+k<t \leq T_{j}$, this equation can always be solved as

$$
\boldsymbol{\beta}^{\prime} \tilde{\boldsymbol{X}}_{t}=\sum_{s=1}^{t-T_{j-1}-k}\left(I_{r}+\boldsymbol{\beta}^{\prime} \boldsymbol{\alpha}\right)^{t-T_{j-1}-k-s} \boldsymbol{\beta}^{\prime} \boldsymbol{\iota} \varepsilon_{T_{j-1}+k+s}+\left(I_{r}+\boldsymbol{\beta}^{\prime} \boldsymbol{\alpha}\right)^{t-T_{j-1}-k} \boldsymbol{\beta}^{\prime} \tilde{\boldsymbol{X}}_{T_{j-1}+k}
$$

and, in the first sub-sample period or $j=1$, the initial value $\boldsymbol{\beta}^{\prime} \tilde{\boldsymbol{X}}_{T_{0}+k}=\boldsymbol{\beta}^{\prime} \tilde{\boldsymbol{X}}_{k}$ can be treated as fixed, so that the process $\boldsymbol{\beta}^{\prime} \tilde{X}_{t}$ for $T_{0}+k<t \leq T_{1}$ becomes uniformly bounded in probability by noting that it equals $\boldsymbol{U}_{t}$ in Assumption A1. Similarly, by iterating over all the other start-up values, the process $\boldsymbol{\beta}^{\prime} \tilde{\boldsymbol{X}}_{t}$ for $T_{j-1}+k<t \leq T_{j}$ and $j=2, \ldots, q$ is also uniformly bounded in probability. Since the number of breaks is finite, $\boldsymbol{\beta}^{\prime} \tilde{\boldsymbol{X}}_{t}$ is uniformly bounded in probability jointly for $1 \leq j \leq q$.

Next, the Granger-Johansen representation (18) implies that, for $1 \leq j \leq q$ and $T_{j-1}+k<t \leq T_{j}$,

$$
\Delta X_{t}=C \varepsilon_{t}+\Delta U_{t}+\tau_{\ell, j} \quad \text { and } \quad \beta^{\prime} X_{t}+\gamma_{j} t=\beta^{\prime} U_{t}+\beta^{\prime} \tau_{c, j},
$$

where

$$
U_{t}=(I-C \Psi) \bar{\beta} \beta^{\prime} \tilde{X}_{t}-C \sum_{i=1}^{k-1} \Gamma_{i} \sum_{\ell=0}^{i-1} \Delta \tilde{X}_{t-\ell}
$$

Since $\beta^{\prime} \tilde{X}_{t}$ is uniformly bounded in probability, it follows that $U_{t}$ is also uniformly bounded in probability. Note that $U_{t}$ is identical throughout all sub-sample periods. The intercepts $\tau_{\ell, j}$ and $\beta^{\prime} \tau_{c, j}$ are eliminated from $\Delta X_{t}$ and $\beta^{\prime} X_{t}+\gamma_{j} t$ respectively, when demeaning them within each sub-sample period. Consequently, we can apply the Law of Large Numbers (A4) in Assumption A1 to

$$
\begin{equation*}
\frac{1}{T_{j}-T_{j-1}-k} \sum_{t=T_{j-1}+k}^{T_{j}}\left\{\binom{V_{t}}{Q_{t}}-\frac{1}{T_{j}-T_{j-1}-k} \sum_{s=T_{j-1}+k}^{T_{j}}\binom{V_{s}}{Q_{s}}\right\}^{\otimes 2} \tag{A8}
\end{equation*}
$$

which, for $1 \leq j \leq q$, converges in probability to a positive definite matrix denoted as

$$
\left(\begin{array}{cccc}
N_{\varepsilon \varepsilon}^{(j)} & N_{\varepsilon x}^{(j)} & N_{\varepsilon \beta}^{(j)} & N_{\varepsilon q}^{(j)} \\
N_{x \varepsilon}^{(j)} & N_{x x}^{(j)} & N_{x \beta}^{(j)} & N_{x q}^{(j)} \\
N_{\beta \varepsilon}^{(j)} & N_{\beta x}^{(j)} & N_{\beta \beta}^{(j)} & N_{\beta q}^{(j)} \\
N_{q \varepsilon}^{(j)} & N_{q x}^{(j)} & N_{q \beta}^{(j)} & N_{q q}^{(j)}
\end{array}\right)
$$

for $N_{\varepsilon \varepsilon}^{(j)}=\Omega$ by Assumption A1. Since both $X_{t-1}$ and $Q_{t}$ consist of the past values of $\varepsilon_{t}$, it follows from (A4) that $N_{\varepsilon \beta}^{(j)}=0$ and $N_{\varepsilon q}^{(j)}=0$ hold. Note that the model equation is

$$
\begin{equation*}
\varepsilon_{t}=\Delta X_{t}-\alpha\left(\beta^{\prime} X_{t-1}+\gamma_{j} t\right)-\Gamma Q_{t}+\mu_{j}, \tag{A9}
\end{equation*}
$$

for $\Gamma=\left(\Gamma_{1}, \ldots, \Gamma_{k-1}\right)$. We can derive three properties from this equation. First, we post-multiply (A9) by $\varepsilon_{t}^{\prime}$ and then exploit $N_{\varepsilon \beta}^{(j)}=0$ and $N_{\varepsilon q}^{(j)}=0$ to find $\Omega=N_{\varepsilon \varepsilon}^{(j)}=N_{x \varepsilon}^{(j)}-0-0=N_{x \varepsilon}^{(j)}$, which is the first property. The next one is

$$
\left(N_{x \beta}^{(j)}, N_{x q}^{(j)}\right)\left(\begin{array}{cc}
N_{\beta \beta}^{(j)} & N_{\beta q}^{(j)}  \tag{A10}\\
N_{q \beta}^{(j)} & N_{q q}^{(j)}
\end{array}\right)^{-1}=(\alpha, \boldsymbol{\Gamma})
$$

where the left-hand side is the limit of the sample regression coefficient for $\Delta X_{t}$ regressed on $\beta^{\prime} X_{t-1}+\gamma_{j} t, Q_{t}$ and an intercept. This property is demonstrated by substituting $\varepsilon_{t}+\alpha\left(\beta^{\prime} X_{t-1}+\right.$ $\left.\gamma_{j} t\right)+\Gamma Q_{t}-\mu_{j}$ from (A9) into $\Delta X_{t}$; we then arrive at the limit result

$$
\left(N_{x \beta}^{(j)}, N_{x q}^{(j)}\right)=\left(N_{\varepsilon \beta}^{(j)}, N_{\varepsilon q}^{(j)}\right)+(\alpha, \Gamma)\left(\begin{array}{cc}
N_{\beta \beta}^{(j)} & N_{\beta q}^{(j)} \\
N_{q \beta}^{(j)} & N_{q q}^{(j)}
\end{array}\right),
$$

from which (A10) follows by noting that $N_{\varepsilon \beta}^{(j)}=0$ and $N_{\varepsilon q}^{(j)}=0$. The third property is

$$
\Omega=N_{x x}^{(j)}-(\alpha, \Gamma)\left(\begin{array}{cc}
N_{\beta \beta}^{(j)} & N_{\beta q}^{(j)}  \tag{A11}\\
N_{q \beta}^{(j)} & N_{q q}^{(j)}
\end{array}\right)\binom{\alpha^{\prime}}{\boldsymbol{\Gamma}^{\prime}}
$$

The left-hand side of (A11) is the limit of the sample product moment of $\varepsilon_{t}$ regressed on $\beta^{\prime} X_{t-1}+\gamma_{j} t, Q_{t}$ and an intercept, due to $N_{\varepsilon \beta}^{(j)}=0$ and $N_{\varepsilon q}^{(j)}=0$. The right-hand side of (A11) is the limit of the sample product moment of $\Delta X_{t}$ regressed on $\beta^{\prime} X_{t-1}+\gamma_{j} t$, $Q_{t}$ and an intercept, where we have exploited the identity (A10).

Now, we return to (A8) and partial out $Q_{t}$ to obtain

$$
\begin{align*}
\frac{1}{T_{j}-T_{j-1}-k} \sum_{t=T_{j-1}+k}^{T_{j}} & \left\{V_{t}-\sum_{s=T_{j-1}+k}^{T_{j}} V_{s} \bar{Q}_{s}^{\prime}\left(\sum_{s=T_{j-1}+k}^{T_{j}} \bar{Q}_{s}^{\otimes 2}\right)^{-1} \bar{Q}_{t}\right\}^{\otimes 2} \\
& \xrightarrow{P}\left(\begin{array}{ccc}
N_{\varepsilon \varepsilon}^{(j)} & N_{\varepsilon x}^{(j)} & N_{\varepsilon \beta}^{(j)} \\
N_{x \varepsilon}^{(j)} & N_{x x}^{(j)} & N_{x \beta}^{(j)} \\
N_{\beta \varepsilon}^{(j)} & N_{\beta x}^{(j)} & N_{\beta \beta}^{(j)}
\end{array}\right)-\left(\begin{array}{c}
N_{\varepsilon q}^{(j)} \\
N_{x \emptyset}^{(j)} \\
N_{\beta q}^{(j)}
\end{array}\right)\left(N_{q q}^{(j)}\right)^{-1}\left(N_{q \varepsilon}^{(j)}, N_{q x}^{(j)}, N_{q \beta}^{(j)}\right), \tag{A12}
\end{align*}
$$

for which we have that $N_{\varepsilon \varepsilon}^{(j)}=N_{\varepsilon x}^{(j)}=\Omega$ while $N_{\varepsilon \beta}^{(j)}=0$ and $N_{\varepsilon q}^{(j)}=0$. Thus, (A12) is reduced to

$$
\left(\begin{array}{ccc}
\Omega & \Omega & 0 \\
\Omega & \Sigma_{x x}^{(j)} & \Sigma_{x \beta}^{(j)} \\
0 & \Sigma_{\beta x}^{(j)} & \Sigma_{\beta \beta}^{(j)}
\end{array}\right)
$$

where

$$
\left(\begin{array}{cc}
\Sigma_{x x}^{(j)} & \Sigma_{x \beta}^{(j)} \\
\Sigma_{\beta x}^{(j)} & \Sigma_{\beta \beta}^{(j)}
\end{array}\right)=\left(\begin{array}{cc}
N_{x x}^{(j)} & N_{x \beta}^{(j)} \\
N_{\beta x}^{(j)} & N_{\beta \beta}^{(j)}
\end{array}\right)-\binom{N_{x q}^{(j)}}{N_{\beta q}^{(j)}}\left(N_{q q}^{(j)}\right)^{-1}\left(N_{q x}^{(j)}, N_{q \beta}^{(j)}\right) .
$$

Furthermore, noting $\sum_{j=1}^{q} \Delta v_{j}=1$ and $\sum_{j=1}^{q}\left(\Delta v_{j}\right) \Omega=\Omega$, we define

$$
\begin{equation*}
\Sigma_{i k}=\sum_{j=1}^{q}\left(\Delta v_{j}\right) \Sigma_{i k}^{(j)} \quad \text { and } \quad N_{l m}=\sum_{j=1}^{q} \Delta v_{j} N_{l m}^{(j)} \tag{A13}
\end{equation*}
$$

for $i, k=x, \beta$ and $l, m=q, x, \beta$. The use of Slutsky's theorem then leads to (A7).
It is left to show that $\Sigma_{x \beta}=\alpha \Sigma_{\beta \beta}$ and $\Sigma_{x x}=\alpha \Sigma_{\beta x}+\Omega$. For the first expression, we apply the identities in (A13) to (A10) so as to obtain

$$
\left(N_{x \beta}, N_{x q}\right)=(\alpha, \boldsymbol{\Gamma})\left(\begin{array}{cc}
N_{\beta \beta} & N_{\beta q}  \tag{A14}\\
N_{q \beta} & N_{q q}
\end{array}\right) .
$$

Taking partitioned inversion in (A14) results in $\alpha=N_{x \beta \cdot q} N_{\beta \beta \cdot q}^{-1}=\Sigma_{x \beta} \Sigma_{\beta \beta}^{-1}$. For the second expression, we apply the identities in (A13) to (A11) to find

$$
\Omega=N_{x x}-(\alpha, \Gamma)\left(\begin{array}{cc}
N_{\beta \beta} & N_{\beta q}  \tag{A15}\\
N_{q \beta} & N_{q q}
\end{array}\right)\binom{\alpha^{\prime}}{\Gamma^{\prime}} .
$$

Inserting (A14) into (A15) and taking its partitioned inversion, we arrive at

$$
\Omega=N_{x x}-\left(N_{x \beta}, N_{x q}\right)\left(\begin{array}{cc}
N_{\beta \beta} & N_{\beta q} \\
N_{q \beta} & N_{q q}
\end{array}\right)^{-1}\binom{N_{\beta x}}{N_{q x}}=N_{x x \cdot q}-N_{x \beta \cdot q} N_{\beta \beta \cdot q}^{-1} N_{\beta x \cdot q}=\Sigma_{x x}-\alpha \Sigma_{\beta x}
$$

by noting from (A14) that $N_{x \beta \cdot q} N_{\beta \beta \cdot q}^{-1}=\alpha$ holds.
Recalling the decomposition $\Delta X_{t}=\left(\Delta Y_{t}^{\prime}, \Delta Z_{t}^{\prime}\right)^{\prime}$, we find the following equivalence in the lower-right submatrix of (A7) in Lemma A3:

$$
\left(\begin{array}{cc}
\Sigma_{x x} & \Sigma_{x \beta} \\
\Sigma_{\beta x} & \Sigma_{\beta \beta}
\end{array}\right)=\left(\begin{array}{ccc}
\Sigma_{y y} & \Sigma_{y z} & \Sigma_{y \beta} \\
\Sigma_{z y} & \Sigma_{z z} & \Sigma_{z \beta} \\
\Sigma_{\beta y} & \Sigma_{\beta z} & \Sigma_{\beta \beta}
\end{array}\right)=\Sigma .
$$

Under the normality assumption for $\varepsilon_{t}$ as in HJNR, we could form the conditional variance of the two elements $\Delta Y_{t}$ and $\beta^{\prime} X_{t-1}+t \gamma_{j}$ given the element $\Delta Z_{t}$. Moving away from normality under Assumptions 2 and 3, we need to consider instead the limit of a product moment matrix consisting of linear combinations of these elements, defined in the following manner:

$$
\left(\begin{array}{cc}
\Sigma_{y y \cdot z} & \Sigma_{y \beta \cdot z}  \tag{A16}\\
\Sigma_{\beta y \cdot z} & \Sigma_{\beta \beta \cdot z}
\end{array}\right)=A^{\prime} \Sigma A=\left(\begin{array}{cc}
\Sigma_{y y} & \Sigma_{y \beta} \\
\Sigma_{\beta y} & \Sigma_{\beta \beta}
\end{array}\right)-\binom{\Sigma_{y z}}{\Sigma_{\beta z}} \Sigma_{z z}^{-1}\left(\Sigma_{z y}, \Sigma_{z \beta}\right),
$$

which appears in (A17) in Lemma A6, and where

$$
A^{\prime}=\left(\begin{array}{ccc}
I & -\Sigma_{y z} \Sigma_{z z}^{-1} & 0 \\
0 & -\Sigma_{\beta z} \Sigma_{z z}^{-1} & I
\end{array}\right)
$$

Let us recall the weak exogeneity condition $\alpha_{z}=0$, which implies

$$
\alpha=\binom{\alpha_{y}}{0} \quad \text { and } \quad \alpha_{\perp}=\left(\begin{array}{cc}
\alpha_{y \perp} & 0 \\
0 & I_{p-m}
\end{array}\right) .
$$

Finally, recall from (4) that the limit variance of innovations in the partial equation equals $\Omega_{y y \cdot z}=\Omega_{y y}-\Omega_{y z} \Omega_{z z}^{-1} \Omega_{z y}$ under Assumption A1. Within each sub-sample period, the setup here is identical to that of HJNR. We therefore obtain the following equation, which adapts Equation (10). 6 in Johansen (1995, Lemma 10.1).

Lemma A4 (HJNR, Lemma 4). Suppose that Assumptions 1 and A1 are satisfied under $\alpha_{z}=0$. Then,

$$
\alpha_{y \perp}\left(\alpha_{y \perp}^{\prime} \Omega_{y y \cdot z} \alpha_{y \perp}\right)^{-1} \alpha_{y \perp}^{\prime}=\Sigma_{y y \cdot z}^{-1}-\Sigma_{y y \cdot z}^{-1} \Sigma_{y \beta \cdot z}\left(\Sigma_{\beta y \cdot z} \Sigma_{y y \cdot z}^{-1} \Sigma_{y \beta \cdot z}\right)^{-1} \Sigma_{\beta y \cdot z} \Sigma_{y y \cdot z}^{-1}
$$

We now explore the limit of the common trends within each sub-sample period. Define

$$
B_{T}^{* \prime}=\left(\begin{array}{cc}
\alpha_{\perp}^{\prime} \Gamma & 0 \\
0 & T^{-1 / 2}
\end{array}\right)\left(\begin{array}{cc}
I_{p} & -\tau_{\ell, j} \\
0 & 1
\end{array}\right) \quad \text { and } \quad X_{t}^{*}=\binom{X_{t-1}}{t}
$$

and the next lemma is a combination of Lemma A. 1 in JMN and Lemma 5 in HJNR.
Lemma A5. Suppose that Assumptions 1 and A1 are satisfied. Consider the ( $p-r+1$ )-dimensional process $T^{-1 / 2} B_{T}^{* \prime} X_{\mathrm{int}(T u)}^{*}$ on $D[0,1]$ endowed with the Skorokhod metric with common distortion across the dimensions. Let $W_{u}$ be a $(p-r)$-dimensional Brownian motion with variance $\Omega$ for $0 \leq u \leq 1$. For $v_{j-1} \leq u<v_{j}$ and $1 \leq j \leq q$, the process $X_{\operatorname{int}(T u)}^{*}$ satisfies

$$
T^{-1 / 2} B_{T}^{* \prime}\left(X_{\mathrm{int}(T u)}^{*}-X_{\mathrm{int}\left(T v_{j-1}\right)}^{*}\right) \xrightarrow{\mathrm{D}}\left\{\begin{array}{c}
\alpha_{\perp}^{\prime}\left(W_{u}-W_{v_{j-1}}\right) \\
u-v_{j-1}
\end{array}\right\} .
$$

The convergence holds jointly for $1 \leq j \leq q$ and $0 \leq u \leq 1$.
Proof of Lemma A5. The Granger-Johansen representation (18) implies that, for $1 \leq j \leq q$ and $T_{j-1}+k<t \leq T_{j}$,

$$
T^{-1 / 2} \alpha_{\perp}^{\prime} \Gamma\left(X_{t}-\tau_{\ell, j} t\right) \approx T^{-1 / 2} \alpha_{\perp}^{\prime} \Gamma\left(C \sum_{s=T_{j-1}+k+1}^{t} \varepsilon_{s}+U_{t}+\tau_{c, j}\right)
$$

Since $\alpha_{\perp}^{\prime} \Gamma C=\alpha_{\perp}^{\prime}$ has full row rank and $U_{t}$ is bounded in probability as shown in the proof of Lemma A3, the random walk component $\alpha_{\perp}^{\prime} \sum_{s=T_{j-1}+k+1}^{t} \varepsilon_{s}$ dominates $\alpha_{\perp}^{\prime} \Gamma U_{t}$. The initial value $\tau_{c, j}$ could be large when $j>1$, but it is eliminated when taking differences $X_{i n t}^{*}(T u)-X_{\operatorname{int}\left(T v_{j-1}\right)}^{*}$. Thus, the first element of $T^{-1 / 2} B_{T}^{* \prime}\left(X_{\text {int }(T u)}^{*}-X_{i n t\left(T v_{j-1}\right)}^{*}\right)$ converges to $\alpha_{\perp}^{\prime}\left(W_{u}-W_{v_{j-1}}\right)$ by the Functional Central Limit Theorem (A5) in Assumption A1. The second element also converges as desired since $T^{-1} \operatorname{int}(T u)$ converges to $u$. As the number of breaks is finite, the convergence holds jointly for $1 \leq j \leq q$.

Decompose $W_{u}=\left(W_{1 u}^{\prime}, W_{2 u}^{\prime}\right)^{\prime}$, in which the dimensions of $W_{1 u}$ and $W_{2 u}$ are $m$ and $p-m$, respectively. Let us recall the notation $X_{t-1}^{\ell}=\left(X_{t-1}^{\prime}, t E_{t}^{\prime}\right)^{\prime}$, which was used in reduced rank regression in Section 3.1. The next lemma establishes the asymptotic theory for the product moment matrices $S_{i j}$ for $i, j=0,1$ defined in (24).

Lemma A6. Suppose that Assumptions 1 and A1 are satisfied under $\alpha_{z}=0$. Define $\beta^{\ell}=\left(\beta^{\prime}, \gamma\right)^{\prime}$, $\tau_{\ell}=\left(\tau_{\ell, 1}, \ldots, \tau_{\ell, q}\right) \in R^{p \times q}$ and

$$
B_{T}^{\ell \prime}=\left(\begin{array}{cc}
\alpha_{\perp}^{\prime} \Gamma & 0 \\
0 & T^{-1 / 2} I_{q}
\end{array}\right)\left(\begin{array}{cc}
I_{p} & -\tau_{\ell} \\
0 & I_{q}
\end{array}\right) .
$$

It then follows that

$$
\begin{align*}
\left(\begin{array}{cc}
S_{00} & S_{01} \beta^{\ell} \\
\beta^{\ell \prime} S_{10} & \beta^{\ell \prime} S_{11} \beta^{\ell}
\end{array}\right) & \xrightarrow{\mathrm{P}}\left(\begin{array}{cc}
\Sigma_{y y \cdot z} & \Sigma_{y \beta \cdot z} \\
\Sigma_{\beta y \cdot z} & \Sigma_{\beta \beta \cdot z}
\end{array}\right)  \tag{A17}\\
T^{-1} B_{T}^{\ell \prime} S_{11} B_{T}^{\ell} & \xrightarrow{\mathrm{D}} \int_{0}^{1} F_{u} F_{u}^{\prime} d u  \tag{A18}\\
B_{T}^{\ell \prime}\left(S_{10}-S_{11} \beta^{\ell} \alpha_{y}^{\prime}\right) & \xrightarrow{\mathrm{D}} \int_{0}^{1} F_{u} d\left(W_{1 u}-\omega W_{2 u}\right),  \tag{A19}\\
B_{T}^{\ell \prime} S_{11} \beta^{\ell} & =\mathrm{OP}_{\mathrm{P}}(1), \tag{A20}
\end{align*}
$$

where

$$
F_{u}=\binom{\alpha_{\perp}^{\prime} W_{u}}{u e_{u}}-\int_{0}^{1}\binom{\alpha_{\perp}^{\prime} W_{s}}{s e_{s}} e_{s}^{\prime} d s\left(\int_{0}^{1} e_{s} e_{s}^{\prime} d s\right)^{-1} e_{u}
$$

Proof of Lemma A6. Recall the decomposition $\Delta X_{t}=\left(\Delta Y_{t}^{\prime}, \Delta Z_{t}^{\prime}\right)^{\prime}$.
For (A17), we start with the left-hand side of (A12) and further partial out $\Delta Z_{t}$ from the process, to which we then apply the Law of Large Numbers (A4) in Assumption A1. Follow the proof of Lemma A3 afterwards, supplemented with the definition of the limit expression (A16), in order to verify (A17).

For (A18), use Lemma A5, the continuous mapping theorem and Johansen (1995, Lemma 10.3).
For (A19), we note $B_{T}^{\ell \prime}\left(S_{10}-S_{11} \beta^{\ell} \alpha_{y}^{\prime}\right)=B_{T}^{\ell \prime} S_{1 \varepsilon}$. The Law of Large Numbers (A4) in Assumption A1 implies $B_{T}^{\ell \prime} S_{1 \varepsilon}=B_{T}^{\ell \prime}(T-k)^{-1} \sum_{t=k+1}^{T} \mathcal{Z}_{1, t-1} \varepsilon_{y \cdot z, t}^{\prime}+o_{\mathrm{P}}(1)$, in which $\mathcal{Z}_{1, t}$ is the demeaned version of $X_{t}^{\ell}$ as defined in (23). By (A3) and the Granger-Johansen representation in Theorem 1, we can replace $B_{T}^{\ell \prime} \mathcal{Z}_{1, t}$ with the demeaned version of $\left(\sum_{s=k+1}^{t} \varepsilon_{s}^{\prime} \alpha_{\perp}, t E_{t}^{\prime}\right)$. The stochastic integral (A6) in Assumption A1 then gives (A19).

For (A20), we follow the strategy used for (A19); see also Johansen (1995, Lemma 10.3).

## Appendix C.3. Proofs of the Theorems in Section 3

Proof of Theorem 2. Follow the proof of Theorem 11.1 in Johansen (1995) by using Lemmas A4 and A6 given above instead of his Lemmas 10.1 and 10.3, and also utilise invariance properties with respect to non-singular linear transformations as in the proof of Theorem 1 in HJNR.

Proof of Theorem 3. The proof presented here is based on Doornik (1998, §9). The asymptotic distribution of the $L R$ test statistic in the partial model in Theorem 2 is rewritten as

$$
\begin{aligned}
& \operatorname{tr}\left\{\int_{0}^{1} d B_{u}^{(m-r)} G_{u}^{\prime}\left(\int_{0}^{1} G_{u} G_{u}^{\prime} d u\right)^{-1} \int_{0}^{1} G_{u} d B_{u}^{(m-r) \prime}\right\} \\
& =\sum_{i=1}^{m-r} \int_{0}^{1} d B_{i, u} G_{u}^{\prime}\left(\int_{0}^{1} G_{u} G_{u}^{\prime} d u\right)^{-1} \int_{0}^{1} G_{u} d B_{i, u}^{\prime}=\sum_{i=1}^{m-r} \mathrm{~T}_{i} .
\end{aligned}
$$

The process $\mathrm{T}_{i}$ for $i=1, \ldots, m-r$ is a function of $B_{i, u}$ and $G_{u}$, both of which are functions of the $(p-r)$-dimensional standard Brownian motion $B_{u}$. Inspection of these functions shows that they are invariant to the relabelling of the coordinates of $B_{u}$, so that $\mathrm{T}_{1}, \ldots, \mathrm{~T}_{m-r}$ are identically distributed and any pairs $\mathrm{T}_{j}, \mathrm{~T}_{k}$ are also identically distributed. Hence,

$$
\begin{aligned}
\mathrm{E}\left(\sum_{i=1}^{m-r} \mathrm{~T}_{i}\right) & =\sum_{i=1}^{m-r} \mathrm{E}\left(\mathrm{~T}_{i}\right)=(m-r) \mathrm{E}\left(\mathrm{~T}_{1}\right), \\
\operatorname{Var}\left(\sum_{i=1}^{m-r} \mathrm{~T}_{i}\right) & =\sum_{j=1}^{m-r} \sum_{k=1}^{m-r} \operatorname{Var}\left(\mathrm{~T}_{j}, \mathrm{~T}_{k}\right)=\sum_{j=1}^{m-r} \operatorname{Var}\left(\mathrm{~T}_{j}\right)+\sum_{j \neq k}^{m-r} \operatorname{Var}\left(\mathrm{~T}_{j}, \mathrm{~T}_{k}\right) \\
& =(m-r) \operatorname{Var}\left(\mathrm{T}_{1}\right)+(m-r)(m-r-1) \operatorname{Cov}\left(\mathrm{T}_{1}, \mathrm{~T}_{2}\right) .
\end{aligned}
$$

In order to relate the moments of the limit distributions of the $L R$ test statistics in the partial and full models, we evaluate the above expressions for $m-r$ in general and for $m-r=p-r$. For the means of the limit distributions, we find

$$
\mathrm{E}\left(\sum_{i=1}^{m-r} \mathrm{~T}_{i}\right)=(m-r) \mathrm{E}\left(\mathrm{~T}_{1}\right) \quad \text { and } \quad \mathrm{E}\left(\sum_{i=1}^{p-r} \mathrm{~T}_{i}\right)=(p-r) \mathrm{E}\left(\mathrm{~T}_{1}\right) .
$$

Solving both equations for $E\left(T_{1}\right)$ and equating the resulting expressions yield

$$
\mathrm{E}\left(\sum_{i=1}^{m-r} \mathrm{~T}_{i}\right)=\left(\frac{m-r}{p-r}\right) \mathrm{E}\left(\sum_{i=1}^{p-r} \mathrm{~T}_{i}\right) .
$$

For their variances, we obtain a set of equations similarly, which are solved for $\operatorname{Var}\left(\mathrm{T}_{1}\right)$ to find the desired expression.

## References

Anderson, Theodore W., and Naoto Kunitomo. 1992. Asymptotic distributions of regression and autoregression coefficients with martingale difference disturbances. Journal of Multivariate Analysis 40: 221-43. [CrossRef]
Bårdsen, Gunnar, Øyvind Eitrheim, Eilev S. Jansen, and Ragnar Nymoen. 2005. The Econometrics of Macroeconomic Modelling. Oxford: Oxford University Press.
Berenguer-Rico, Vanessa, and Bent Nielsen. 2017. Marked and Weighted Empirical Processes of Residuals with Applications to Robust Regressions. Discussion Paper 841. Oxford: Department of Economics, University of Oxford.
Berenguer-Rico, Vanessa, and Ines Wilms. 2018. White Heteroscedasticity Testing in Robust Regressions. Discussion Paper 853. Oxford: University of Oxford, Department of Economics.
Billingsley, Patrick. 1968. Convergence of Probability Measures. New York: John Wiley \& Sons.
Boswijk, H. Peter. 1992. Cointegration, Identification and Exogeneity, 3rd ed. Tinbergen Institute Research Series. Amsterdam: Thesis Publishers, volume 37.

Boswijk, H. Peter 1995. Efficient inference on cointegration parameters in structural error correction models. Journal of Econometrics 69: 133-58. [CrossRef]
Boswijk, H. Peter, Giuseppe Cavaliere, Anders Rahbek, and A. M. Robert Taylor. 2016. Inference on co-integration parameters in heteroskedastic vector autoregressions. Journal of Econometrics 192: 64-85. [CrossRef]
Boswijk, H. Peter, and Jurgen A. Doornik. 2005. Distribution approximations for cointegration tests with stationary exogenous regressors. Journal of Applied Econometrics 20: 797-810. [CrossRef]
Brown, B. M. 1971. Martingale central limit theorems. Annals of Mathematical Statistics 42: 59-66. [CrossRef]
Cavaliere, Giuseppe, Luca De Angelis, and Luca Fanelli. 2018. Co-integration rank determination in partial systems using information criteria. Oxford Bulletin of Economics and Statistics 80: 65-89. [CrossRef]
Cavaliere, Giuseppe, Anders Rahbek, and A. M. Robert Taylor. 2010. Cointegration rank testing under conditional heteroskedasticity. Econometric Theory 26: 1719-60. [CrossRef]
Cavaliere, Giuseppe, Anders Rahbek, and A. M. Robert Taylor. 2012. Bootstrap determination of the co-integration rank in vector autoregressive models. Econometrica 80: 1721-40.
Chan, Ngai Hang, and Ching Zong Wei. 1988. Limiting distributions of least squares estimates of unstable autoregressive processes. Annals of Statistics 16: 367-401. [CrossRef]
Doornik, Jurgen A. 1998. Approximations to the asymptotic distribution of cointegration tests. Journal of Economic Surveys 12: 573-93. [CrossRef]
Doornik, Jurgen A. 2003. Asymptotic Tables for cointegration Tests Based on the GAmma-Distribution Approximation. Available online: www.doornik.com/research/coigamma_tables.pdf (accessed on 4 October 2019).

Doornik, Jurgen A. 2013. An Object-Oriented Matrix Programming Language-OX7. London: Timberlake.
Doornik, Jurgen A., and Henrik Hansen. 2008. An omnibus test for univariate and multivariate normality. Oxford Bulletin of Economics and Statistics 70: 927-39. [CrossRef]
Doornik, Jurgen A., and David F. Hendry. 2013. PcGive 14. London: Timberlake, volume 2.
Doornik, Jurgen A., David F. Hendry, and Bent Nielsen. 1998. Inference in cointegrating models: UK M1 revisited. Journal of Economic Surveys 12: 533-72. [CrossRef]
Engle, Robert F. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica 50: 987-1108. [CrossRef]
Engle, Robert F., David F. Hendry, and Jean-Francois Richard. 1983. Exogeneity. Econometrica 51: 277-304. [CrossRef]
Ericsson, Neil R., and James G. MacKinnon. 2002. Distributions of error correction tests for cointegration. Econometrics Journal 5: 285-318. [CrossRef]
Godfrey, Leslie G. 1978. Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. Econometrica 46: 1293-301. [CrossRef]
Hall, P., and C. C. Heyde. 1980. Martingale Limit Theory and its Applications. San Diego: Academic Press.
Hannan, E. J., and C. C. Heyde. 1972. On limit theorems for quadratic functions of discrete time series. Annals of Mathematical Statistics 43: 2058-66. [CrossRef]
Hansen, Bruce E. 1992. Convergence to stochastic integrals for dependent heterogeneous processes. Econometric Theory 8: 489-500. [CrossRef]
Hansen, Peter Reinhard 2005. Granger's representation theorem: A closed-form expression for I(1) processes. Econometrics Journal 8: 23-38.
Harbo, Ingrid, Søren Johansen, Bent Nielsen, and Anders Rahbek. 1998. Asymptotic inference on cointegrating rank in partial systems. Journal of Business and Economic Statistics 16: 388-99.
Harvey, Andrew, and Stephen Thiele. 2017. Co-Integration and Control: Assessing the Impact of Events Using Time Series Data. Working Paper 1731, University of Cambridge, Faculty of Economics, Cambridge, UK.
Hendry, David F., and Michael Massmann. 2007. Co-breaking: Recent advances and a synopsis of the literature. Journal of Business \& Economic Statistics 25: 33-51.
Inoue, Atsushi. 1999. Tests of cointegrating rank with a trend-break. Journal of Econometrics 90: 215-37. [CrossRef]
Jakubowski, Adam, Jean Ménin, and Gilles Pages. 1989. Convergence en loi des suites d'integrales stochastiques sur l'espace $D^{1}$ de skorokhod. Probability Theory and Related Fields 81: 111-37. [CrossRef]
Johansen, Søren. 1988. Statistical analysis of cointegration vectors. Journal of Economic Dynamics \& Control 12: 231-54.

Johansen, Søren. 1992a. Cointegration in partial systems and the efficiency of single-equation analysis. Journal of Econometrics 52: 389-402. [CrossRef]
Johansen, Søren. 1992b. Testing weak exogeneity and the order of cointegration in the UK money demand data. Journal of Policy Modeling 14: 313-34. [CrossRef]
Johansen, Søren. 1995. Likelihood Based Inference on Cointegration in the Vector Autoregressive Model. Oxford: Oxford University Press.
Johansen, Søren, Rocco Mosconi, and Bent Nielsen. 2000. Cointegration analysis in the presence of structural breaks in the deterministic trend. Econometrics Journal 3: 216-49. [CrossRef]
Juselius, Katarina 2006. The Cointegrated VAR Model. Oxford: Oxford University Press.
Kurita, Takamitsu. 2011. Local power of likelihood-based tests for cointegrating rank: Comparative analysis of full and partial systems. Journal of Time Series Analysis 32: 672-79. [CrossRef]
Kurita, Takamitsu, and Bent Nielsen. 2009. Cointegrated vector autoregressive models with adjusted short-run dynamics. Quantitative and Qualitative Analysis in Social Sciences 3: 43-77.
Kurita, Takamitsu, Heino Bohn Nielsen, and Anders Rahbek. 2011. An I(2) cointegration model with piecewise linear trends. Econometrics Journal 14: 131-55. [CrossRef]
Kurtz, Thomas G., and Philip E. Protter. 1991. Weak limit theorems for stochastic integrals and stochastic differential equations. Annals of Probability 19: 1035-70. [CrossRef]
Kurtz, Thomas G., and Philip E. Protter. 1996. Weak convergence of stochastic integrals and differential equations. In Probabilistic Models for Nonlinear Partial Differential Equations. Edited by Denis Talay and Luciano Tubaro. Lecture Notes in Mathematics. Berlin: Springer, volume 1627, pp. 197-285.
Lai, Tze Leung, and Ching Zong Wei. 1982. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. Annals of Statistics 10: 154-66. [CrossRef]
Lai, T. L., and C. Z. Wei. 1985. Asymptotic properties of multivariate weighted sums with applications to stochastic regression in linear dynamic systems. In Multivariate Analysis-VI. Edited by P. R. Krishnaiah. Amsterdam: Elsevier, pp. 375-93.
MacKinnon, James G., Alfred A. Haug, and Leo Michelis. 1999. Numerical distributions of likelihood ratio tests of cointegration. Journal of Applied Econometrics 14: 563-77. [CrossRef]
Nielsen, B. 1997. Bartlett correction of the unit root test in autoregressive models. Biometrika 84: 500-4. [CrossRef]
Nielsen, Bent. 2005. Strong consistency results for least squares estimators in general vector autoregressions with deterministic terms. Econometric Theory 21: 534-61. [CrossRef]
Nielsen, Bent. 2006. Order determination in general vector autoregressions. In Time Series and Related Topics: In Memory of Ching-Zong Wei. Edited by Hwai-Chung Ho, Ching-Kang Ing and Tze Leung Lai. Lecture Notes-Monograph Series. Beachwood: Institute of Mathematical Statistics, vol. 52, pp. 93-112.
Paruolo, Paolo. 1997. Asymptotic inference on the moving average impact matrix in cointegrated I(1) var systems. Econometric Theory 13: 79-118. [CrossRef]
Paruolo, Paolo, and Anders Rahbek. 1999. Weak exogeneity in I(2) VAR systems. Journal of Econometrics 93: 281-308. [CrossRef]
Perron, Pierre. 1989. The great crash, the oil price shock, and the unit root hypothesis. Econometrica 57: 1361-401. Erratum in volume 61, 248-49. [CrossRef]
Pesaran, M. Hashem, Yongcheol Shin, and Richard J. Smith. 2000. Structural analysis of vector error correction models with exogenous I(1) variables. Journal of Econometrics 97: 293-343. [CrossRef]
Rahbek, Anders, and Rocco Mosconi. 1999. Cointegration rank inference with stationary regressors in VAR models. Econometrics Journal 2: 76-91. [CrossRef]
Schreiber, Vanessa. 2015. Cointegration Analysis of the Exchange Rate and Bilateral Trade between the United Kingdom and Germany. M.Phil. thesis, University of Oxford, Oxford, UK.
Seo, Byeongseon. 1998. Statistical inference on cointegration rank in error correction models with stationary covariates. Journal of Econometrics 85: 339-85. [CrossRef]
White, Halbert. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48: 817-38. [CrossRef]
© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http:/ /creativecommons.org/licenses/by/4.0/).

## Article

# Modeling I(2) Processes Using Vector Autoregressions Where the Lag Length Increases with the Sample Size 

Yuanyuan Li and Dietmar Bauer *<br>Faculty of Business Administration and Economics, Bielefeld University, Universitätsstrasse 25, D-33615 Bielefeld, Germany; yuan_yuan.li@uni-bielefeld.de<br>* Correspondence: Dietmar.Bauer@uni-bielefeld.de

Received: 4 May 2018; Accepted: 8 September 2020; Published: 17 September 2020


#### Abstract

In this paper the theory on the estimation of vector autoregressive (VAR) models for I(2) processes is extended to the case of long VAR approximation of more general processes. Hereby the order of the autoregression is allowed to tend to infinity at a certain rate depending on the sample size. We deal with unrestricted OLS estimators (in the model formulated in levels as well as in vector error correction form) as well as with two stage estimation (2SI2) in the vector error correction model (VECM) formulation. Our main results are analogous to the $I(1)$ case: We show that the long VAR approximation leads to consistent estimates of the long and short run dynamics. Furthermore, tests on the autoregressive coefficients follow standard asymptotics. The pseudo likelihood ratio tests on the cointegrating ranks (using the Gaussian likelihood) used in the 2SI2 algorithm show under the null hypothesis the same distributions as in the case of data generating processes following finite order VARs. The same holds true for the asymptotic distribution of the long run dynamics both in the unrestricted VECM estimation and the reduced rank regression in the 2SI2 algorithm. Building on these results we show that if the data is generated by an invertible VARMA process, the VAR approximation can be used in order to derive a consistent initial estimator for subsequent pseudo likelihood optimization in the VARMA model.


Keywords: vector autoregressions; vector error correction model; integrated processes of order two

## 1. Introduction

Many macroeconomic variables have been found to exhibit trend-like behaviour that can be modelled by using vector autoregressions (VARs). Katarina Juselius (2006) states that empirical modelling led to the development of $\mathrm{I}(1)$ and $\mathrm{I}(2)$ models since certain features of the datasets considered required including first and second differences in order to obtain stationary time series. Additionally cointegrating relations were found in the corresponding analyses. Similar findings have reoccurred numerous times in the literature for example related to money demand Johansen (1992b); Juselius (1994), inflation Banerjee et al. (2001); Georgoutsos and Kouretas (2004), interest rates and real exchange rates Johansen et al. (2007); Juselius and Assenmacher (2017); Juselius and Stillwagon (2018); Stillwagon (2018) to mention only a few sources.

The predominant methodological approach to model integration and cointegration in the $\mathrm{I}(1)$ and the I(2) case in the vector autoregressive (VAR) framework has been established mainly by Søren Johansen and Katarina Juselius together with a number of coauthors (see the lists of references in Johansen (1995); Juselius (2006) for details) building on vector error correction models (see Engle and Granger (1987) for early comments on the history of using error correction models for co-integrated processes). Extending the main ideas for cointegration modeling for the I(1) setting Johansen (1997) see, e.g., Johansen (1992a) suggested a representation for the $I(2)$ case. Johansen (1997) established asymptotic distributions for the suggested two step $I(2)$ estimator (2SI2) as an approximation to pseudo maximum likelihood
estimation involving numerical optimization. Asymptotics for the corresponding likelihood ratio tests has been developed in Paruolo $(1994,1996)$, its asymptotic equivalence to pseudo likelihood (using the Gaussian distribution) optimization (and hence in a certain sense statistical efficiency) is shown in Paruolo (2000). However, Nielsen and Rahbek (2007) shows that in finite samples the likelihood ratio test has size advantages. The testing of restrictions on the parameters has been investigated by Boswijk and Doornik (2004); Boswijk and Paruolo (2017); Johansen and Lütkepohl (2005). Due to the implicit vector error correction (VECM) modeling, deterministic terms in the VECM produce complex deterministic terms in the solutions processes. In the I(2) context Nielsen and Rahbek (2007); Paruolo (1994, 2006); Rahbek et al. (1999); Kurita et al. (2011) discuss the impacts of deterministic terms.

As the VECM representation includes the representation of reduced rank matrices by a product of two matrices, identification conditions are of particular importance, see Juselius (2006); Mosconi and Paruolo (2013, 2017). In this context also weak exogeneity has been studied Kurita (2012); Paruolo and Rahbek (1999).

The main idea underlying the VECM approach for estimating VAR models in the I(2) context is to reparameterize the problem such that integration and cointegration properties relate to the rank of two matrices. Assuming the data generating process to be a VAR of known finite order, the rank of matrices can be tested using (pseudo) likelihood ratio tests.

Sometimes the assumption of known order is not justified. For example it is known that a subset of variables that are generated using a finite order VAR cannot be described by a finite order VAR, but instead requires a vector autoregressive moving average (VARMA) model. However, the class of VARs provides flexibility in the sense that a VAR of infinite order can represent a large set of linear dynamical systems including all invertible VARMA systems. For stationary processes Berk (1974) and Lewis and Reinsel (1985) show that by letting the order of the VAR tend to infinity at a suitable function of the sample size, consistent estimation of the underlying transfer function can be achieved for data generating processes that can be described by a $\operatorname{VAR}(\infty)$ subject to mild assumptions on the summability of the VAR coefficients. Additionally Lewis and Reinsel (1985) also establishes asymptotic normality (in a very specific sense) of linear combinations of the estimated autoregressive coefficients. Hannan and Deistler (1988) make the concepts operational by showing that in the case of a VARMA process generating the dataset the required rate of letting the order tend to infinity can be estimated using BIC model selection.

In the case of $\mathrm{I}(1)$ processes the estimation theory for long VAR approximations to VARMA processes has been extended based on the techniques in the stationary case of Lewis and Reinsel in a series of papers by Saikkonen and coauthors Saikkonen (1991, 1992); Lütkepohl and Saikkonen (1997); Saikkonen and Lütkepohl (1996); Saikkonen and Luukkonen (1997). Additionally also the Johansen framework of rank restricted estimation in the VECM model has been extended to the long VAR approximations by Saikkonen and Luukkonen (1997). Bauer and Wagner (2004) provide extensions to the multi frequency $I(1)$ case where unit roots may occur at the seasonal frequencies.

For the $I(2)$ case no such extensions are currently known. This is the research gap this paper tries to fill: First we establish consistency and asymptotic normality of estimated autoregressive coefficients (in the sense of Lewis and Reinsel) for unrestricted ordinary least squares (OLS) estimation in the VECM representation. This can be used in order to derive Wald type tests of linear restrictions on the autoregressive parameters. Secondly, we extend the rank restricted regression techniques in the $I(2)$ case to the long VAR approximations showing that the asymptotics (for estimated cointegrating relations, likelihood ratio tests and the two step estimation procedures) are identical in the case of long VAR approximations and VARs of finite known order. Third, we show that if the data generating process is an invertible VARMA process the long VAR system estimator can be used in order to obtain consistent initial estimators for subsequent pseudo likelihood maximization in the VARMA model class. In all results we limit ourselves to the case of no deterministic terms being included in the VECM representation. The inclusion of deterministic terms requires changing the test distribution, compare the theory contained for example in Rahbek et al. (1999).

The paper is organized as follows: In the next section the data generating process and the main assumptions are described. Section 3 then provides the results for the unrestricted estimation. Section 4 deals with rank restricted regression in the 2SI2 procedure, while Section 5 investigates the initial guess in the VARMA setting for subsequent pseudo likelihood maximization. Finally Section 6 concludes the paper. Proofs are relegated to an appendix.

Throughout the paper we will use the notation introduced by Johansen (1997): For a matrix $C \in \mathbb{R}^{p \times s}, s<p$, of full column rank we use the notation $\bar{C}=C\left(C^{\prime} C\right)^{-1}$. Furthermore, $C_{\perp}$ denotes a full column rank matrix of dimension $p \times(p-s)$ such that $C_{\perp}^{\prime} C=0$. Whenever this notation is used the particular choice of $C_{\perp}$ is not of importance. For a matrix $\stackrel{\rightharpoonup}{C}=\left(C_{i, j}\right) \in \mathbb{R}^{p \times s}$ we let $\|C\|$ denote the Frobenius norm $\|C\|=\sqrt{\sum_{i=1}^{p} \sum_{j=1}^{S} C_{i, j}^{2}}$.

## 2. Data Generating Process and Assumptions

In this paper we use the following assumptions on the data generating process:
Assumption 1 (DGP). The process $\left(y_{t}\right)_{t \in \mathbb{Z}}, y_{t} \in \mathbb{R}^{p}$, is generated from the difference equation for $t \in \mathbb{Z}$ :

$$
\begin{equation*}
\Delta^{2} y_{t}=\alpha \beta^{\prime} y_{t-1}+\Gamma \Delta y_{t-1}+\sum_{j=1}^{\infty} \Pi_{j} \Delta^{2} y_{t-j}+\varepsilon_{t} \tag{1}
\end{equation*}
$$

where $\alpha, \beta \in \mathbb{R}^{p \times r}, 0 \leq r<p$ are full column rank matrices, $\Delta=(1-L)$ with $L$ denoting the backward shift operator such that $L\left(y_{t}\right)_{t \in \mathbb{Z}}=\left(y_{t-1}\right)_{t \in \mathbb{Z}}$. The matrix function $A(z)=(1-z)^{2} I_{p}-\alpha \beta^{\prime} z-\Gamma z(1-z)-$ $\sum_{j=1}^{\infty} \Pi_{j}(1-z)^{2} z^{j}$ fulfills the special marginal stability condition that

$$
\begin{equation*}
|A(z)|=0 \quad \text { implies that } \quad|z|>1 \quad \text { or } \quad z=1 . \tag{2}
\end{equation*}
$$

Furthermore, there exists a real $\delta>0$ such that the power series defining $A(z)$ converges absolutely for $|z|<1+\delta$. Define $\beta_{2}=\beta_{\perp} \eta_{\perp}, \alpha_{2}=\alpha_{\perp} \zeta_{\perp}$ where $\alpha_{\perp}^{\prime} \Gamma \beta_{\perp}=\zeta \eta^{\prime}, \eta, \zeta \in \mathbb{R}^{(p-r) \times s}$ are of full column rank $s<p-r$. Then it is assumed that the matrix

$$
\begin{equation*}
\alpha_{2}^{\prime}\left(I_{p}+\Gamma \bar{\beta} \bar{\alpha}^{\prime} \Gamma-\sum_{j=1}^{\infty} \Pi_{j}\right) \beta_{2} \tag{3}
\end{equation*}
$$

is nonsingular.
Furthermore, the process $\left(\varepsilon_{t}\right)_{t \in \mathbb{Z}}$ denotes independent identically distributed (iid) white noise with mean zero and variance $\Sigma_{\epsilon}>0$.

It is well known that the conditions (2) and (3) are necessary and sufficient for the existence of solutions to the difference equation that are $\mathrm{I}(2)$ processes, see for example Johansen (1992a). Moreover, note that the assumption of absolute convergence of $A(z)$ for $|z|<1+\delta$ implies that $\sum_{j=0}^{\infty} j^{k}\left\|\Pi_{j}\right\|<\infty$ for every $k \in \mathbb{N}$. In particular $\sum_{j=0}^{\infty} j^{2}\left\|\Pi_{j}\right\|<\infty$ follows as will be used frequently below.

Every vector autoregressive function $A(z)$ corresponding to the autoregression $A(L) y_{t}=\varepsilon_{t}$, that fulfills Assumption 1, allows a representation as $A(z)=(1-z)^{2} I_{p}-\alpha \beta^{\prime} z-\Gamma z(1-z)-$ $\sum_{j=1}^{\infty} \Pi_{j}(1-z)^{2} z^{j}=\tilde{g}(z) \tilde{B}(z), \tilde{B}(z)=(1-z)^{2} I_{p}-\tilde{\Pi} z-\tilde{\Gamma} z(1-z), \tilde{g}(z)=I_{p}+\sum_{j=1}^{\infty} G_{j} z^{j}$. This can be seen as follows:

$$
\begin{aligned}
\varepsilon_{t} & =A(L) y_{t}=\left(A(1)-\dot{A}(1) \Delta+A^{*}(L) \Delta^{2}\right) y_{t}=\left(A(1)-\dot{A}(1) \Delta+A^{*}(L) \Delta^{2}\right) \mathcal{B} \mathcal{B}^{\prime} y_{t} \\
& =\left([-\alpha, 0,0]+[\alpha, 0,0] \Delta-\Gamma \mathcal{B} \Delta+A^{*}(L) \mathcal{B} \Delta^{2}\right) \mathcal{B}^{\prime} y_{t} \\
& =\left(\left[-\alpha,-\Gamma \beta_{1},-\Gamma \beta_{2}\right]+\left[\alpha-\Gamma \beta, A_{1}^{*}(L), A_{2}^{*}(L)\right] \Delta+\left[A_{0}^{*}(L), 0,0\right] \Delta^{2}\right)\left(\begin{array}{c}
\beta^{\prime} \\
\beta_{1}^{\prime} \Delta \\
\beta_{2}^{\prime} \Delta
\end{array}\right) y_{t} \\
& =\left(\left[-\alpha,-\Gamma \beta_{1},-\Gamma \beta_{2}+\alpha \bar{\alpha}^{\prime} \Gamma \beta_{2}\right]+\left[\alpha-\Gamma \beta, A_{1}^{*}(L), \tilde{A}_{2}^{*}(L)\right] \Delta+\left[A_{0}^{*}(L), 0,-A_{0}^{*}(L) \bar{\alpha}^{\prime} \Gamma \beta_{2}\right] \Delta^{2}\right)\left(\begin{array}{c}
\beta^{\prime}+\bar{\alpha}^{\prime} \Gamma \beta_{2} \beta_{2}^{\prime} \Delta \\
\beta_{1}^{\prime} \Delta \\
\beta_{2}^{\prime} \Delta
\end{array}\right) y_{t} \\
& =\left(\left[-\alpha,-\Gamma \beta_{1}, \tilde{A}_{2}^{*}(L)\right]+\left[\alpha-\Gamma \beta, A_{1}^{*}(L),-A_{0}^{*}(L) \bar{\alpha}^{\prime} \Gamma \beta_{2}\right] \Delta+\left[A_{0}^{*}(L), 0,0\right] \Delta^{2}\right)\left(\begin{array}{c}
\beta^{\prime}+\bar{\alpha}^{\prime} \Gamma \beta_{2} \beta_{2}^{\prime} \Delta \\
\beta_{1}^{\prime} \Delta \\
\beta_{2}^{\prime} \Delta^{2}
\end{array}\right) y_{t} \\
& =g(L) B(L) y_{t}
\end{aligned}
$$

where $\mathcal{B}=\left[\beta, \beta_{1}, \beta_{2}\right], \beta_{1}=\beta_{\perp} \eta$, is without restriction of generality assumed to be an orthonormal matrix, $A^{*}(L) \mathcal{B}=\left[A_{0}^{*}(L), A_{1}^{*}(L), A_{2}^{*}(L)\right], A(1)=-\alpha \beta^{\prime}, \dot{A}(1)=-\alpha \beta^{\prime}+\Gamma$ and where we use that

$$
\Gamma \beta_{2}-\alpha \bar{\alpha}^{\prime} \Gamma \beta_{2}=\left(I_{p}-\alpha \bar{\alpha}^{\prime}\right) \Gamma \beta_{2}=\bar{\alpha}_{\perp} \alpha_{\perp}^{\prime} \Gamma \beta_{\perp} \eta_{\perp}=0 .
$$

Here

$$
B(L)=\left(\begin{array}{c}
\beta^{\prime}+\bar{\alpha}^{\prime} \Gamma \beta_{2} \beta_{2}^{\prime} \Delta \\
\beta_{1}^{\prime} \Delta \\
\beta_{2}^{\prime} \Delta^{2}
\end{array}\right)
$$

In this representation

$$
g(1)=\left[-\alpha,-\Gamma \beta_{1}, \tilde{A}_{2}^{*}(1)\right]
$$

is nonsingular due to assumption (3). Furthermore, $g(z)=\sum_{j=0}^{\infty} G_{j} z^{j}$ is a transfer function with $\sum_{j=0}^{\infty}\left\|G_{j}\right\| j^{2}<\infty$ since $\sum_{j=1}^{\infty}\left\|\Pi_{j}\right\| j^{2}<\infty$ and thus the same holds for the power series coefficients $A^{*}(L)$. Since $|B(z)| \neq 0, z \neq 1$ it follows that $|g(z)| \neq 0,|z| \leq 1$. Therefore

$$
\begin{equation*}
B(L) y_{t}=u_{t}, \quad g(L) u_{t}=\varepsilon_{t} \tag{4}
\end{equation*}
$$

is a VAR process. Note, however, that $g(0)=G_{0} \neq I_{p}$ in general. This constitutes a triangular representation of the process denoting $y_{1, t}=\beta^{\prime} y_{t} \in \mathbb{R}^{p_{1}}, y_{2, t}=\beta_{1}^{\prime} y_{t} \in \mathbb{R}^{p_{2}}, y_{3, t}=\beta_{2}^{\prime} y_{t} \in \mathbb{R}^{p_{3}}$ such that

$$
\begin{aligned}
y_{1, t} & =-\bar{\alpha}^{\prime} \Gamma \beta_{2} \Delta y_{3, t}+u_{1, t}=A \Delta y_{3, t}+u_{1, t} \quad A: p_{1} \times p_{3} \\
\Delta y_{2, t} & =u_{2, t} \\
\Delta^{2} y_{3, t} & =u_{3, t}
\end{aligned}
$$

where $u_{t}=\left[u_{1, t}^{\prime}, u_{2, t}^{\prime}, u_{3, t}^{\prime}\right]^{\prime}$ has a $\operatorname{VAR}(\infty)$ representation. Furthermore, defining

$$
\begin{aligned}
& \tilde{B}(L)=\mathcal{B}\left(\begin{array}{ccc}
I_{p_{1}} & 0 & -\bar{\alpha}^{\prime} \Gamma \beta_{2} \\
0 & I_{p_{2}} & 0 \\
0 & 0 & I_{p_{3}}
\end{array}\right) B(L)=\Delta^{2} I_{p}+\beta \beta^{\prime} L+\left(\beta \beta^{\prime}+\beta \bar{\alpha}^{\prime} \Gamma \beta_{2} \beta_{2}^{\prime}+\beta_{1} \beta_{1}^{\prime}\right) L \Delta, \\
& \tilde{g}(L)=g(L)\left(\mathcal{B}\left(\begin{array}{ccc}
I_{p_{1}} & 0 & -\bar{\alpha}^{\prime} \Gamma \beta_{2} \\
0 & I_{p_{2}} & 0 \\
0 & 0 & I_{p_{3}}
\end{array}\right)\right)^{-1}
\end{aligned}
$$

we obtain $A(L)=g(L) B(L)=\tilde{g}(L) \tilde{B}(L)$ such that

$$
\tilde{B}(L) y_{t}=\Delta^{2} y_{t}+\tilde{\Pi} y_{t-1}+\tilde{\Gamma} \Delta y_{t-1}=v_{t}, \quad \tilde{g}(L) v_{t}=\varepsilon_{t}
$$

is another representation of the process $\left(y_{t}\right)_{t \in \mathbb{Z}}$ with $\tilde{B}(0)=I_{p}$. It follows that the triangular representation can be seen as a special case where one has partial information on the matrices $\beta, \beta_{1}, \beta_{2}$. For estimation the VECM representation is approximated using a finite order $h$ :

$$
\Delta^{2} y_{t}=\Phi y_{t-1}+\Psi \Delta y_{t-1}+\sum_{j=1}^{h-2} \Pi_{j} \Delta^{2} y_{t-j}+e_{t}
$$

where $e_{t}=\varepsilon_{t}+e_{1 t}, e_{1 t}=\sum_{j=h-1}^{\infty} \Pi_{j} \Delta^{2} y_{t-j}$. As in the VECM representation the dimensions of $\beta, \beta_{1}, \beta_{2}$ are linked to the rank of the matrices $\Phi$ and $\alpha_{\perp}^{\prime} \Psi \beta_{\perp}$. Restricting these matrices to be of particular rank is simpler than imposing the equivalent restrictions in the VAR(h) representation directly.

In the following we will first investigate the unrestricted ordinary least squares estimator in the VECM representation without taking rank restrictions into account. In the second step the 2SI2 procedure as presented in Paruolo (2000) for imposing the two rank restrictions in two steps is investigated.

For both procedures the selection of the order $h$ is of importance. In this respect the following assumption will be used:

Assumption 2 (Lag order $h$ ). The order $h$ is chosen subject to the following restrictions:

1. $h=o\left(T^{1 / 5}\right)$.
2. $\quad T^{1 / 2} \sum_{j=h+1}^{\infty}\left\|\Pi_{j}\right\| \rightarrow 0$ as $T, h \rightarrow \infty$.

This condition defines an upper bound for the order which is usually directly assured during order selection using for example information criteria. The upper bound is smaller than the usual rate $T^{1 / 3}$ for technical reasons. The stronger bound is not needed for all results. However, the implications for practical applications are minor as for example in the range $1 \leq T \leq 950$ we have $2.5 T^{1 / 5}>T^{1 / 3}$. The second condition of Assumption 2 implies a lower bound for the increase of $h$ as a function of the sample size. Clearly $\sum_{j=h+1}^{\infty}\left\|\Pi_{j}\right\| \rightarrow 0$ for $h \rightarrow \infty$. The bound implies that for $h=h(T)$ this convergence needs to be fast enough such that $T^{1 / 2} \sum_{j=h(T)+1}^{\infty}\left\|\Pi_{j}\right\|$ still converges to zero. The lower bound depends on the underlying true parameters. For invertible VARMA processes - which can be seen as the leading case $-\left\|\Pi_{j}\right\| \leq C \rho_{0}^{j}$ for some $0 \leq \rho_{0}<1$. Hannan and Deistler (1988) show that for an invertible stationary VARMA process the lower bound (in this case proportional to $\log T$ ) can be achieved asymptotically by using BIC as the order selection procedure. Thus in this case also the stronger condition $\left(h=o\left(T^{1 / 5}\right)\right)$ is satisfied. Bauer and Wagner (2004) extend this result to the multi frequency $I(1)$ setting. For the $I(2)$ case no analogous result is known, although the developments of Bauer and Wagner (2004) suggest that a similar result holds also there. This is left for future research.

Therefore the difference between the 'usual' rates and the ones assumed above are deemed to be of minor practical consequences. Thus we are not explicit in the main text as to which results hold true under the less restrictive set of results and which do not. In the appendix, we will comment on this point, however.

## 3. Unrestricted Estimation

In this section the results of Lewis and Reinsel (1985) and Saikkonen and Lütkepohl (1996) are extended to the $\mathrm{I}(2)$ case. To simplify notation define $\left\langle a_{t}, b_{t}\right\rangle=\sum_{t=h+1}^{T} a_{t} b_{t}^{\prime}$ for sequences $a_{t}, b_{t}, t=$ $1, \ldots, T .{ }^{1}$ Then the unrestricted least squares estimator in the finite VECM model uses the regressor vector $Z_{t, h}=\left[y_{t-1}^{\prime}, \Delta y_{t-1}^{\prime}, \Delta^{2} y_{t-1}^{\prime}, \ldots, \Delta^{2} y_{t-h+2}^{\prime}\right]^{\prime} \in \mathbb{R}^{p h}$. The corresponding ordinary least squares estimator is given as

[^25]\[

$$
\begin{aligned}
{\left[\hat{\Phi}, \hat{\Psi}, \hat{\Pi}_{1}, \ldots, \hat{\Pi}_{h-2}\right] } & =\left[\left\langle\Delta^{2} y_{t}, y_{t-1}\right\rangle,\left\langle\Delta^{2} y_{t}, \Delta y_{t-1}\right\rangle,\left\langle\Delta^{2} y_{t}, \Delta^{2} y_{t-1}\right\rangle, \ldots,\left\langle\Delta^{2} y_{t}, \Delta^{2} y_{t-h+2}\right\rangle\right]\left\langle Z_{t, h}, Z_{t, h}\right\rangle^{-1} \\
& =\left\langle\Delta^{2} y_{t}, Z_{t, h}\right\rangle\left\langle Z_{t, h}, Z_{t, h}\right\rangle^{-1} .
\end{aligned}
$$
\]

The noise covariance is estimated from the residuals as usual as

$$
\begin{equation*}
\hat{\Sigma}_{\epsilon}=N^{-1}\left\langle\hat{e}_{t}, \hat{e}_{t}\right\rangle, \quad \hat{e}_{t}=\Delta^{2} y_{t}-\hat{\Phi} y_{t-1}-\hat{\Psi} \Delta y_{t-1}-\sum_{j=1}^{h-2} \hat{\Pi}_{j} \Delta^{2} y_{t-j} \tag{5}
\end{equation*}
$$

where $N=T-h$ denotes the effective sample size.

### 3.1. Estimation in the Triangular VECM Representation

As typical for the cointegration framework, analysis is easier in the triangular representation which separates stationary components from $I(1)$ and $I(2)$ processes: Let $y_{t}=\left[y_{1, t}^{\prime}, y_{2, t}^{\prime}, y_{3, t}^{\prime}\right]^{\prime} \in \mathbb{R}^{p}$ where $y_{i, t} \in \mathbb{R}^{p_{i}}$ is such that

$$
\begin{aligned}
y_{1, t} & =A \Delta y_{3, t}+u_{1, t}, \\
\Delta y_{2, t} & =u_{2, t}, \\
\Delta^{2} y_{3, t} & =u_{3, t}
\end{aligned}
$$

where $u_{t}=\left[u_{1, t}^{\prime}, u_{2, t}^{\prime}, u_{3, t}^{\prime}\right]^{\prime}$ has a $\operatorname{VAR}(\infty)$ representation $g(L) u_{t}=\varepsilon_{t}$ where

$$
g(0)=\left(\begin{array}{ccc}
I & 0 & A \\
0 & I & 0 \\
0 & 0 & I
\end{array}\right)
$$

Note, however, that using the triangular representation implies that the matrix $B(L)$ is known up the value of the matrix $A$. For applications this is the case only seldom.

Thus letting $g(z)=g(1)+g^{*}(z) \Delta$ we obtain

$$
\begin{aligned}
\varepsilon_{t} & =g(L)\left(\begin{array}{c}
y_{1, t}-A \Delta y_{3, t} \\
\Delta y_{2, t} \\
\Delta^{2} y_{3, t}
\end{array}\right)=g(L)\left(\begin{array}{c}
\Delta^{2} y_{1, t}+\Delta y_{1, t-1}+y_{1, t-1}-A \Delta^{2} y_{3, t}-A \Delta y_{3, t-1} \\
\Delta^{2} y_{2, t}+\Delta y_{2, t-1} \\
\Delta^{2} y_{3, t}
\end{array}\right) \\
& =g(L)\left(\begin{array}{ccc}
I & 0 & -A \\
0 & I & 0 \\
0 & 0 & I
\end{array}\right) \Delta^{2} y_{t}+g(L)\left(\begin{array}{c}
y_{1, t-1} \\
0 \\
0
\end{array}\right)+g(L)\left(\begin{array}{c}
\Delta y_{1, t-1}-A \Delta y_{3, t-1} \\
\Delta y_{2, t-1} \\
0
\end{array}\right) \\
& =\tilde{g}(L) \Delta^{2} y_{t}+\left[g(1)+g^{*}(L) \Delta\right]\left(\begin{array}{c}
y_{1, t-1} \\
0 \\
0
\end{array}\right)+g(1)\left(\begin{array}{c}
\Delta y_{1, t-1}-A \Delta y_{3, t-1} \\
\Delta y_{2, t-1} \\
0
\end{array}\right) \\
& \left.=\pi(L) \Delta^{2} y_{t}+g(1)\left(\begin{array}{c}
y_{1, t-1} \\
0 \\
0
\end{array}\right)+\left[\begin{array}{ll}
\mathcal{G}_{1}+\mathcal{G}_{1}^{*} & \mathcal{G}_{2}
\end{array}\right)-\mathcal{G}_{1} A\right]\left(\begin{array}{c}
\Delta y_{1, t-1} \\
\Delta y_{2, t-1} \\
\Delta y_{3, t-1}
\end{array}\right) \\
& =\pi(L) \Delta^{2} y_{t}+\left[\begin{array}{lll}
\mathcal{G}_{1} & 0 & 0
\end{array}\right] y_{t-1}+\left[\begin{array}{lll}
\mathcal{G}_{1}+\mathcal{G}_{1}^{*} & \mathcal{G}_{2} & -\mathcal{G}_{1} A
\end{array}\right] \Delta y_{t-1}
\end{aligned}
$$

with $\pi(L)=I_{p}-\sum_{j=1}^{\infty} \Pi_{j} L^{j}$ leads to the corresponding VECM representation:

$$
\Delta^{2} y_{t}=\Phi y_{t-1}+\Psi \Delta y_{t-1}+\sum_{j=1}^{\infty} \Pi_{j} \Delta^{2} y_{t-j}+\varepsilon_{t}
$$

Here $\mathcal{G}:=g(1)=\sum_{j=0}^{\infty} G_{j}=\left[\mathcal{G}_{1}, \mathcal{G}_{2}, \mathcal{G}_{3}\right]$, where $\mathcal{G}_{i}$ is $p \times p_{i}$ for $i=1,2,3$ : Similarly, $\mathcal{G}^{*}:=g^{*}(1)=$ $-\sum_{j=0}^{\infty} j G_{j}=\left[\mathcal{G}_{1}^{*}, \mathcal{G}_{2}^{*}, \mathcal{G}_{3}^{*}\right]$, where $\mathcal{G}_{i}^{*}$ is $p \times p_{i}$ for $i=1,2,3$. The sums exists since $\sum_{j=1}^{\infty}\left\|G_{j}\right\| j^{2}<\infty$ by assumption. Similarly, we partition $\Phi, \Psi$ and $\Pi_{j}$ into $\left[\Phi_{1}, \Phi_{2}, \Phi_{3}\right],\left[\Psi_{1}, \Psi_{2}, \Psi_{3}\right]$ and $\left[\Pi_{j 1}, \Pi_{j 2}, \Pi_{j 3}\right]$, respectively. The analogous partitioning is used for estimates.

Then $\Phi=-\left[\mathcal{G}_{1}, 0,0\right], \Psi=\left[-\mathcal{G}_{1}^{*}-\mathcal{G}_{1},-\mathcal{G}_{2}, \mathcal{G}_{1} A\right]$. Therefore $\Psi_{3}=-\Phi_{1} A$. Note that in this notation the $\mathrm{I}(2)$ components on the right hand side are $y_{t-1,3}$, the $\mathrm{I}(1)$ components are $y_{t-1,1}, y_{t-1,2}, \Delta y_{t-1,3}$, where $y_{t-1,1}-A \Delta y_{t-1,3}$ is stationary. Thus in order to separate regressors of different integration orders in the proof (as is usually done in the literature) we use a transformation using the unknown matrix $A$ such that the regressor $y_{t-1,1}$ is replaced by $y_{t-1,1}-A \Delta y_{t-1,3}$. Consequently the estimate $\hat{\Psi}_{3}$ of $\Psi_{3}$ is replaced by the estimate $\hat{\Theta}=\hat{\Psi}_{3}+\hat{\Phi}_{1} A$ of $\Theta=\Psi_{3}+\Phi_{1} A=0$.

Based on the estimates $\hat{\Psi}$ and $\hat{\Phi}$ then $A$ can be estimated as

$$
\begin{equation*}
\hat{A}=-\left(\hat{\Phi}_{1}^{\prime} \hat{\Sigma}_{\epsilon}^{-1} \hat{\Phi}_{1}\right)^{-1} \hat{\Phi}_{1}^{\prime} \hat{\Sigma}_{\epsilon}^{-1} \hat{\Psi}_{3} \tag{6}
\end{equation*}
$$

Here the insertion of $\hat{\Sigma}_{\epsilon}^{-1}$ appears somewhat arbitrary. A motivation for this choice in the $\mathrm{I}(1)$ case can be found in Saikkonen (1992) equation (12). However, any other positive definite matrix could be used as well. Currently there is no knowledge on the optimality of the choice suggested above.

In the asymptotic distribution of the estimation error Brownian motions occur relating to the process $\left(u_{t}\right)_{t \in \mathbb{Z}}$ : Under Assumption 1 we have

$$
\frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor r T\rfloor} u_{t} \Rightarrow B(r)=\left[B_{1}(r)^{\prime}, B_{c}(r)^{\prime}\right]^{\prime}=\left[B_{1}(r)^{\prime}, B_{2}(r)^{\prime}, B_{3}(r)\right]^{\prime}
$$

where $B(r), 0 \leq r \leq 1$, denotes a Brownian motion with corresponding variance

$$
\Omega=\left[\begin{array}{c|c}
\Omega_{11} & \Omega_{1 c} \\
\hline \Omega_{c 1} & \Omega_{c c}
\end{array}\right]=\left[\begin{array}{l|ll}
\Omega_{11} & \Omega_{12} & \Omega_{13} \\
\hline \Omega_{21} & \Omega_{22} & \Omega_{23} \\
\Omega_{31} & \Omega_{32} & \Omega_{33}
\end{array}\right]=g(1)^{-1} \Sigma_{\epsilon}\left(g(1)^{\prime}\right)^{-1}
$$

where $B_{1 . c}(r)=B_{1}(r)-\Omega_{1 c} \Omega_{c c}^{-1} B_{c}(r)$ is a $p_{1}$-dimensional Brownian motion, which is independent of $B_{c}(r)$, with covariance

$$
\Omega_{1 . c}=\Omega_{11}-\Omega_{1 c} \Omega_{c c}^{-1} \Omega_{c 1} .
$$

An estimator of $\Omega_{1 . c}$ is given by ${ }^{2}$

$$
\begin{equation*}
\hat{\Omega}_{1 . c}=\left(\hat{\Phi}_{1}^{\prime} \hat{\Sigma}_{\epsilon}^{-1} \hat{\Phi}_{1}\right)^{-1} \tag{7}
\end{equation*}
$$

With these definitions we can state our first result of the paper (which is proved in Appendix B):
Theorem 1. Under Assumptions 1 and 2 for the triangular VECM representation we have:
(A) Consistency:
(i) $\quad \hat{\Phi} \xrightarrow{p} \Phi ; \quad$ (ii) $\quad \hat{\Sigma}_{\epsilon} \xrightarrow{p} \Sigma_{\epsilon} ; \quad$ (iii) $\quad \hat{\Omega}_{1 . c} \xrightarrow{p} \Omega_{1 . c} ; \quad$ (iv) $\quad \hat{\Psi} \xrightarrow{p} \Psi ; \quad$ (v) $\quad \hat{\Theta} \xrightarrow{p} 0 ; \quad$ (vi) $\quad \hat{A} \xrightarrow{p} A$.
(B) Asymptotic distribution of coefficients to nonstationary regressors: Under Assumptions 1 and 2 we have ( $N=T-h$ ):

$$
\begin{equation*}
\text { (i) }\left[N \hat{\Phi}_{2}, N \hat{\Theta}, N^{2} \hat{\Phi}_{3}\right] \xrightarrow{d} g(1) \int_{0}^{1} d B F^{\prime}\left(\int_{0}^{1} F F^{\prime}\right)^{-1}, \quad \text { (ii) } \quad N(\hat{A}-A) \xrightarrow{d} \int_{0}^{1} d B_{1 . c} L^{\prime}\left(\int_{0}^{1} L L^{\prime}\right)^{-1} \tag{8}
\end{equation*}
$$

[^26]where $F(u)=\left[\begin{array}{c}B_{c}(u) \\ \int_{0}^{u} B_{3}(v) d v\end{array}\right], F_{a}(u)=\left[\begin{array}{c}B_{2}(u) \\ \int_{0}^{u} B_{3}(v) d v\end{array}\right]$ and $L(u)=B_{3}(u)-\int_{0}^{1} B_{3} F_{a}^{\prime}\left(\int_{0}^{1} F_{a} F_{a}^{\prime}\right)^{-1} F_{a}(u)$.
(C) Asymptotic distribution of coefficients to stationary regressors: Let $L_{h}$ be a sequence of $\left(p^{2}(h-2)+\right.$ $\left.p\left(2 p_{1}+p_{2}\right)\right) \times J$ matrices such that $L_{h}^{\prime}\left(\Gamma_{E C M}^{-1} \otimes \Sigma_{\epsilon}\right) L_{h} \rightarrow M>0$ where $\Gamma_{E C M}=\mathbb{E}\left(X_{t} X_{t}^{\prime}\right)$ with $X_{t}:=$ $\left[u_{1, t-1}^{\prime}, \Delta y_{1, t-1}^{\prime}, \Delta y_{2, t-1}^{\prime}, \Delta^{2} y_{t-1}^{\prime}, \ldots, \Delta^{2} y_{t-h+2}^{\prime}\right]^{\prime}$.

Let

$$
\underline{\Pi}=\left[\begin{array}{llllll}
\Phi_{1} & \Psi_{1} & \Psi_{2} & \Pi_{1} & \ldots & \Pi_{h-2}
\end{array}\right] .
$$

Then

$$
N^{\frac{1}{2}} L_{h}^{\prime} \operatorname{vec}(\hat{\Pi}-\underline{\Pi}) \xrightarrow{d} N(0, M) .
$$

(D) Asymptotic distribution on Wald type tests: Finally letting

$$
\hat{\Gamma}_{E C M}=N^{-1}\left(\left\langle\tilde{X}_{t}, \tilde{X}_{t}\right\rangle-\left\langle\tilde{X}_{t}, \Delta y_{3, t-1}\right\rangle\left\langle\Delta y_{3, t-1}, \Delta y_{3, t-1}\right\rangle^{-1}\left\langle\Delta y_{3, t-1}, \tilde{X}_{t}\right\rangle\right)
$$

where $\tilde{X}_{t}=\left[y_{1, t-1}^{\prime}, \Delta y_{1, t-1}^{\prime}, \Delta y_{2, t-1}^{\prime}, \Delta^{2} y_{t-1}^{\prime}, \ldots, \Delta^{2} y_{t-h+2}^{\prime}\right]^{\prime}$, the Wald test for the null hypothesis $H_{0}$ : $L_{h}^{\prime} \operatorname{vec}(\underline{\Pi})=l_{h}$ is given by

$$
\hat{\lambda}_{\text {Wald }}=N\left(L_{h}^{\prime} \operatorname{vec}(\hat{\Pi})-l_{h}\right)^{\prime}\left(L_{h}^{\prime}\left(\hat{\Gamma}_{E C M}^{-1} \otimes \hat{\Sigma}_{\epsilon}\right) L_{h}\right)^{-1}\left(L_{h}^{\prime} \operatorname{vec}(\hat{\Pi})-l_{h}\right) .
$$

Then if $L_{h}$ is such that $L_{h}^{\prime}\left(\Gamma_{E C M}^{-1} \otimes \Sigma_{\epsilon}\right) L_{h} \rightarrow M>0$, under the null hypothesis $\hat{\lambda}_{\text {Wald }} \xrightarrow{d} \chi^{2}(J)$.
The theorem provides the asymptotic distributions of the OLS estimates in the triangular system. Note that in this somewhat special case the properties of the regressor components (stationary or not) are known such that for each entry the convergence speed is known. Correspondingly the definition of the regressor vector $\tilde{X}_{t}$ involves only lags of $y_{t}$ but omits all nonstationary regressors except the ones cointegrated with $\Delta y_{3, t-1}$.

The assumptions on $L_{h}$ are more restrictive than needed. Lewis and Reinsel (1985) and Saikkonen and Lütkepohl (1996) only require that $L_{h}$ has full column rank when deriving the normalized convergence to normal distribution with unit variance as the limit for

$$
N^{\frac{1}{2}}\left(L_{h}^{\prime}\left(\Gamma_{E C M}^{-1} \otimes \Sigma_{\epsilon}\right) L_{h}\right)^{-1 / 2} L_{h}^{\prime} \operatorname{vec}(\hat{\Pi}-\underline{\Pi}) .
$$

Similar arguments could be used here.

### 3.2. Estimation in the General VECM Representation

The previous section dealt with the special case that a triangular representation is used and hence knowledge on the matrices $\left[\beta, \beta_{1}, \beta_{2}\right]$ is given. This section provides a result for the general case, which, however, is limited to the coefficients to the stationary components. Since a general process generated according to Assumption 1 can be rewritten into a triangular representation using the knowledge of $\left[\beta, \beta_{1}, \beta_{2}\right]$, some asymptotic properties of the unrestricted OLS estimators can be derived from Theorem 1 for the general case (which is proved in Appendix C):

Theorem 2. Let the regressor vector $Z_{t, h}=\left[y_{t-1}^{\prime}, \Delta y_{t-1}^{\prime}, \Delta^{2} y_{t-1}^{\prime}, \ldots, \Delta^{2} y_{t-h+2}^{\prime}\right]$ and define

$$
\underline{\Lambda}=\left[\begin{array}{lllll}
\Phi & \Psi & \Pi_{1} & \ldots & \Pi_{h-2}
\end{array}\right], \quad \tilde{\Lambda}=\left\langle\Delta^{2} y_{t}, Z_{t, h}\right\rangle\left\langle Z_{t, h}, Z_{t, h}\right\rangle^{-1}, \quad \tilde{\Gamma}_{E C M}=N^{-1}\left\langle Z_{t, h}, Z_{t, h}\right\rangle
$$

Then under Assumptions 1 and 2 it follows that $\tilde{\Lambda}-\underline{\Lambda}=o_{P}(1)$.

Furthermore, let $L_{h} \in \mathbb{R}^{p^{2}(h+2) \times J}$ be such that $L_{h}^{\prime}\left(\tilde{\Gamma}_{E C M}^{-1} \otimes \Sigma_{\epsilon}\right) L_{h} \rightarrow M>0$. Then

$$
N^{\frac{1}{2}} L_{h}^{\prime} \operatorname{vec}(\tilde{\Lambda}-\underline{\Lambda}) \xrightarrow{d} N(0, M)
$$

Beside consistency the theorem implies that linear combination of OLS estimators show asymptotic normality and hence standard inference, if the asymptotic variance is nonsingular. One application of such results consists in the so called 'surplus lag' formulation in the context of Granger causality testing, see Bauer and Maynard (2012); Dolado and Lütkepohl (1996).

Finally note that this section does not contain results with regard to the cointegrating rank or the cointegrating space. The theorem above merely allows to test coefficients corresponding to stationary regressors. Therefore the usage is limited to somewhat special situations like the surplus-lag causality tests. However, it is also relevant for impulse response analysis, compare Inoue and Kilian (2020).

## 4. Rank Restricted Regression

The previous sections show that for the estimators discussed in that sections full inference on all coefficients is only possible when information on the matrices $\beta, \beta_{1}$ and $\beta_{2}$ exists. The dimensions of the matrices relate to the ranks of the matrices $\Phi=\alpha \beta^{\prime}$ and, conditional on this, to the rank of $\bar{\alpha}_{\perp}^{\prime} \Psi \bar{\beta}_{\perp}$. The two rank restrictions make estimation and specification more complex than in the I(1) case.

Johansen (1995) provides the two-step approach 2SI2 that can be used for estimation and specification of the two integer valued parameters $p_{1}$ and $p_{2}$. Paruolo and Rahbek (1999) extend the 2SI2 procedure suggested in section 8 of Johansen (1997). Paruolo (2000) shows that this 2SI2 procedure achieves the same asymptotic distribution as pseudo maximum likelihood estimation which could be performed subsequent to 2SI2 estimation. This makes the procedure attractive from a practical point of view. In this section we show that these approaches extend naturally to the long VAR case. The main focus here lies on the derivation of the asymptotic properties of the rank tests.

Recall the long VAR approximation given as

$$
\begin{equation*}
\Delta^{2} y_{t}=\Phi y_{t-1}+\Psi \Delta y_{t-1}+\sum_{j=1}^{h-2} \Pi_{j} \Delta^{2} y_{t-j}+e_{t} \tag{9}
\end{equation*}
$$

where $\Phi=\alpha \beta^{\prime}$ has reduced rank $r<p$ and $\bar{\alpha}_{\perp}^{\prime} \Psi \bar{\beta}_{\perp}=\zeta \eta^{\prime}$ has reduced rank $s<p-r$. In this notation the 2SI2 procedure works as follows: In the first step the rank constraint on $\bar{\alpha}_{\perp}^{\prime} \Psi \bar{\beta}_{\perp}$ is neglected estimating $\alpha$ and $\beta$ by using reduced-rank regression (RRR). Then in the second step the reduced rank of $\bar{\alpha}_{\perp}^{\prime} \Psi \bar{\beta}_{\perp}$ is imposed using RRR in a transformed equation.

In more detail using the Johansen notation we denote with $R_{0 t}, R_{1 t}$ and $R_{2 t}$ the residuals of regressing $\Delta^{2} y_{t}, \Delta y_{t-1}$ and $y_{t-1}$ on $\Delta^{2} y_{t-1}, \ldots, \Delta^{2} y_{t-h+2}$, respectively; then we can rewrite (9) as

$$
\begin{equation*}
R_{0 t}=\alpha \beta^{\prime} R_{2 t}+\Psi R_{1 t}+\tilde{e}_{t} \tag{10}
\end{equation*}
$$

Concentrating out $R_{1 t}$ and denoting the residuals as $R_{0.1 t}$ and $R_{2.1 t}$ we obtain with $S_{i j .1}=$ $\left\langle R_{i t}, R_{j t}\right\rangle-\left\langle R_{i t}, R_{1 t}\right\rangle\left\langle R_{1 t}, R_{1 t}\right\rangle^{-1}\left\langle R_{1 t}, R_{j t}\right\rangle$ the solution to the RRR problem from solving the eigenvalue problem

$$
\begin{equation*}
\left|\lambda S_{22.1}-S_{20.1} S_{00.1}^{-1} S_{02.1}\right|=0, \tag{11}
\end{equation*}
$$

with solutions $1>\hat{\lambda}_{1} \geq \ldots \geq \hat{\lambda}_{p}>0$ ordered with decreasing size and corresponding vectors $V=\left(v_{1}, \ldots, v_{p}\right)$. Then as usual the trace statistic of testing the model $H_{r}$ with $\operatorname{rank}(\Phi) \leq r, r<p$, in the model $H_{p}$ with $\operatorname{rank}(\Phi) \leq p$, is given as

$$
\begin{equation*}
Q_{r}=-2 \log Q\left(H_{r} \mid H_{p}\right)=-T \sum_{i=r+1}^{p} \log \left(1-\hat{\lambda}_{i}\right) \tag{12}
\end{equation*}
$$

The optimizers for $\alpha, \beta$ are given by

$$
\begin{equation*}
\hat{\beta}=\left(v_{1}, \ldots, v_{r}\right), \quad \hat{\alpha}=S_{02.1} \hat{\beta}, \quad \hat{\Sigma}_{\epsilon}=S_{00.1}-\hat{\alpha} \hat{\alpha}^{\prime} . \tag{13}
\end{equation*}
$$

In the second step, given $\alpha$ and $\beta$ known, we can obtain by multiplying (10) by $\bar{\alpha}_{\perp}^{\prime}$ that

$$
\begin{equation*}
\bar{\alpha}_{\perp}^{\prime} R_{0 t}=\bar{\alpha}_{\perp}^{\prime} \Psi\left(\bar{\beta}_{\perp} \beta_{\perp}^{\prime}+\bar{\beta} \beta^{\prime}\right) R_{1 t}+\bar{\alpha}_{\perp}^{\prime} \tilde{e}_{t}=\zeta \eta^{\prime}\left(\beta_{\perp}^{\prime} R_{1 t}\right)+C\left(\beta^{\prime} R_{1 t}\right)+\bar{\alpha}_{\perp}^{\prime} \tilde{e}_{t} . \tag{14}
\end{equation*}
$$

Note that $\beta^{\prime} R_{1 t}$ is stationary. Thus concentrating out $C$ and denoting the residuals as $R_{\bar{\alpha}_{\perp}, \beta, t}$ and $R_{\beta_{\perp} \cdot \beta, t}$, respectively, we can define $S_{a b \cdot \beta}:=\left\langle R_{a . \beta, t}, R_{b . \beta, t}\right\rangle$, for $a, b=\bar{\alpha}_{\perp}$ or $\beta_{\perp}$. Then the likelihood ratio test of the model $H_{r, s}$ with $\operatorname{rank}\left(\zeta \eta^{\prime}\right) \leq s, s<p-r$ in the model $H_{r}^{0}$ with $\operatorname{rank}\left(\bar{\alpha}_{\perp}^{\prime} \Psi \bar{\beta}_{\perp}\right)=p-r$ is given by

$$
\begin{equation*}
Q_{r, s}=-2 \log Q\left(H_{r, s} \mid H_{r}^{0}\right)=-T \sum_{i=s+1}^{p-r} \log \left(1-\hat{\rho}_{i}\right) \tag{15}
\end{equation*}
$$

where $1>\hat{\rho}_{1} \geq \ldots \geq \hat{\rho}_{p-r}>0$ are the solutions of the eigenvalue problem

$$
\begin{equation*}
\mid \rho S_{\beta_{\perp} \beta_{\perp} \cdot \beta}-S_{\beta_{\perp} \underline{\underline{\alpha}}_{\perp} \cdot \beta} S_{\underline{\alpha}_{\perp}}^{-1} \bar{\alpha}_{\perp} \cdot \beta \text { 除 } S_{\perp} \cdot \beta \mid=0, \tag{16}
\end{equation*}
$$

and the corresponding eigenvectors are $W=\left(w_{1}, \ldots, w_{p-r}\right)$. Estimators of $\zeta$ and $\eta$ are given by

$$
\begin{equation*}
\hat{\eta}=\left(w_{1}, \ldots, w_{s}\right), \quad \hat{\zeta}=S_{\bar{\alpha}_{\perp} \beta_{\perp} \cdot \beta} \hat{\eta} \tag{17}
\end{equation*}
$$

For the 2SI2 procedure in this second step the first step estimates $\hat{\alpha}$ and $\hat{\beta}$ are used in place of the unknown true quantities. Then we obtain the following analogon to the results in the finite order VAR framework (the proof is given in Appendix D):

Theorem 3. Let the data be generated according to Assumption 1 and let the VAR order fulfill Assumption 2. Then the following asymptotic results hold:
(A) The asymptotic distribution of the likelihood ratio statistic $Q_{r}$ under the null hypothesis $H_{r}$ is given by

$$
\begin{equation*}
Q_{r} \xrightarrow{d} \operatorname{tr}\left\{\int_{0}^{1} d W_{+} F_{+}^{\prime}\left(\int_{0}^{1} F_{+} F_{+}^{\prime} d u\right)^{-1} \int_{0}^{1} F_{+} d W_{+}^{\prime}\right\} . \tag{18}
\end{equation*}
$$

where $W_{\dagger}=\left(\alpha_{\perp}^{\prime} \Sigma_{\epsilon} \alpha_{\perp}\right)^{-1 / 2} \alpha_{\perp}^{\prime} W, F_{a}(u)=\left[\begin{array}{c}B_{2}(u) \\ \int_{0}^{u} B_{3}(v) d v\end{array}\right]$ and $F_{\dagger}(u)=F_{a}(u)-$ $\int_{0}^{1} F_{a} B_{3}^{\prime}\left(\int_{0}^{1} B_{3} B_{3}^{\prime}\right)^{-1} B_{3}(u)$. This is identical to the distribution achieved in the finite VAR case.
(B) The asymptotic distribution of the likelihood ratio statistic $Q_{r, s}$ under the null hypothesis $H_{r, s}$ is given by

$$
\begin{equation*}
Q_{r, s} \stackrel{d}{\rightarrow} \operatorname{tr}\left\{\int_{0}^{1} d W_{2}^{\prime} B_{3}^{\prime}\left(\int_{0}^{1} B_{3} B_{3}^{\prime} d u\right)^{-1} \int_{0}^{1} B_{3} d W_{2}^{\prime}\right\} \tag{19}
\end{equation*}
$$

where $W_{2}(u)=\left(\alpha_{2}^{\prime} \Sigma_{\epsilon} \alpha_{2}\right)^{-1 / 2} \alpha_{2}^{\prime} W(u)$.
(C) The asymptotic distribution of the test statistic $S_{r, s}=Q_{r}+Q_{r, s}$ under the null hypothesis $H_{r, s}$ is given by

$$
\begin{equation*}
S_{r, s} \xrightarrow{d} \operatorname{tr}\left\{\int_{0}^{1} d W_{+} F_{+}^{\prime}\left(\int_{0}^{1} F_{+} F_{+}^{\prime} d u\right)^{-1} \int_{0}^{1} F_{+} d W_{+}^{\prime}\right\}+\operatorname{tr}\left\{\int_{0}^{1} d W_{2} B_{3}^{\prime}\left(\int_{0}^{1} B_{3} B_{3}^{\prime} d u\right)^{-1} \int_{0}^{1} B_{3} d W_{2}^{\prime}\right\} . \tag{20}
\end{equation*}
$$

(D) Using suitable normalizations all estimators are consistent: $\hat{\alpha}\left(c_{\alpha}^{\prime} \hat{\alpha}\right)^{-1} \xrightarrow{p} \alpha, \hat{\beta}\left(c_{\beta}^{\prime} \hat{\beta}\right)^{-1} \xrightarrow{p} \beta, \hat{\zeta}\left(c_{\zeta}^{\prime} \hat{\zeta}\right)^{-1} \xrightarrow{p}$ $\zeta, \hat{\eta}\left(c_{\eta}^{\prime} \hat{\eta}\right)^{-1} \xrightarrow{p} \eta, \hat{\Psi} \xrightarrow{p} \Psi, \hat{\Pi}_{j} \xrightarrow{p} \Pi_{j}$ where for example $c_{\alpha}^{\prime} \alpha=I_{r}$.
(E) The asymptotic distributions of the coefficients to the nonstationary regressors are identical to the ones in the
finite order VAR case stated in Paruolo (2000). The asymptotic distribution of the coefficients $\hat{\Pi}_{j}$ are identical to the ones in Theorem 1.

The main message of the theorem is that the 2SI2 procedure shows the same asymptotic properties including the rank tests as in the finite order VAR case. As usual also restricting the coefficients for the non-stationary regressors does not influence the asymptotics for the coefficients corresponding to the stationary regressors.

Note that Paruolo (2000) shows that in the finite VAR case 2SI2 estimates have the same asymptotic distribution as pseudo maximum likelihood ( pML ) estimates maximizing the Gaussian likelihood. The first order conditions for the pML estimates of the coefficients to the non-stationary regressors provided in the first display on p. 548 in Paruolo (2000) depend on the data only via the matrices $S_{i j}$ defined above. These matrices depend on the lag length of the VECM only via the concentration step. The proof of our Theorem 3 shows that these terms have the same asymptotic distributions for the finite order VAR and the long VAR. Theorem 4.3 of Paruolo (2000) shows that the asymptotic distribution of the coefficients due to stationary regressors does not depend on the distribution of the coefficients corresponding to the non-stationary regressors as long as they are estimated super-consistently. Thus our results imply that also in the long VAR case the asymptotic distribution of all estimates for the 2SI2 and the pML approach is identical.

## 5. Initial Guess for VARMA Estimation

One usage of long VAR approximations is as preliminary estimate for VARMA model estimation. Hannan and Kavalieris (1986) provide properties of such an approach in the stationary case, Lütkepohl and Claessen (1997) extend the procedure to the $\mathrm{I}(1)$ case. Here we extend this idea to the $I(2)$ case.

The goal is to provide a consistent initial guess for the estimation of a VARMA model for I(2) processes. In this respect we assume the following data generating process:

Assumption 3 (VARMA dgp). The process $\left(y_{t}\right)_{t \in \mathbb{Z}}$ is generated as the solution to the state space equations

$$
\begin{equation*}
y_{t}=C x_{t}+\varepsilon_{t}, \quad x_{t+1}=A x_{t}+B \varepsilon_{t} \tag{21}
\end{equation*}
$$

where $\left(\varepsilon_{t}\right)_{t \in \mathbb{Z}}$ denotes white noise subject to the same assumptions as in Assumption 1.
Here $x_{t} \in \mathbb{R}^{n}$ is the unobserved state process. The system $(A, B, C)$ is assumed to be minimal and in the canonical form of Bauer and Wagner (2012), that is

$$
A=\left[\begin{array}{cccc}
I_{c} & I_{c} & 0 & 0 \\
0 & I_{c} & 0 & 0 \\
0 & 0 & I_{d} & 0 \\
0 & 0 & 0 & A_{\bullet}
\end{array}\right], \quad B=\left[\begin{array}{l}
B_{1} \\
B_{2} \\
B_{3} \\
B_{\bullet}
\end{array}\right], \quad C=\left[\begin{array}{llll}
C_{1} & C_{2} & C_{3} & C_{\bullet}
\end{array}\right],
$$

where $\left|\lambda_{\max }\left(A_{\bullet}\right)\right|<1$ (the matrix $A_{\bullet}$ is stable), $C_{1}^{\prime} C_{1}=I_{c}, C_{3}^{\prime} C_{3}=I_{d}, C_{1}^{\prime} C_{3}=0, C_{1}^{\prime} C_{2}=0, C_{2}^{\prime} C_{3}=0$. Furthermore, the system is strictly minimum-phase, that is $\rho_{0}=\left|\lambda_{\max }(A-B C)\right|<1$. Finally the matrix $\bar{A}=A-B C$ is nonsingular.

At time $t=0$ the state $x_{0}=\left[x_{0, u}^{\prime}, x_{\bullet}^{\prime}\right]^{\prime}, x_{0, u} \in \mathbb{R}^{2 c+d}$, is such that $x_{0, u}$ is deterministic and $x_{0, \bullet}=$ $\sum_{j=1}^{\infty} A_{\bullet}^{j-1} B_{\bullet} \varepsilon_{-j}$ denotes the stationary solution to the stable part of the system.

In this situation it follows that $\left(y_{t}\right)_{t \in \mathbb{Z}}$ is an $\mathrm{I}(2)$ process in the definition of Bauer and Wagner (2012), that is its second difference is a stationary VARMA process. The integers $c$ and $d$ are connected to the integers $p_{1}, p_{2}, p_{3}$ via $c=p_{3}, d=p_{2}$ such that $p_{1}=p-c-d$. It can furthermore be shown that a process generated using Assumption 3 possesses a $\operatorname{VAR}(h)$ approximation:

$$
y_{t}+\sum_{j=1}^{h} A_{j} y_{t-j}=\varepsilon_{t}+C(A-B C)^{h} x_{t-h}
$$

where $A_{j}=-C(A-B C)^{j-1} B,\left\|A_{j}\right\| \leq \mu \rho^{j}\left(0 \leq \rho_{0}<\rho<1\right)$ converges to zero exponentially fast for $j \rightarrow \infty$ due to the strict minimum-phase condition. Letting $h \rightarrow \infty$ then implies the existence of a $\operatorname{VAR}(\infty)$ representation. It follows that for such systems $A(z)$ converges absolutely for $|z|<\rho^{-1}$ where $1<\rho^{-1}$.

From the autoregressive representation the VECM representation can be obtained:

$$
a(z)=I_{p}+\sum_{j=1}^{\infty} A_{j} z^{j}=I_{p}-\sum_{j=1}^{\infty} C \bar{A}^{j-1} B z^{j}=(1-z)^{2} I_{p}-\Phi z-\Psi z(1-z)-(1-z)^{2} \sum_{j=1}^{\infty} \Pi_{j} z^{j}
$$

where $\bar{A}=A-B C$ such that

$$
\Delta^{2} y_{t}=\Phi y_{t-1}+\Psi \Delta y_{t-1}+\sum_{j=1}^{\infty} \Pi_{j} \Delta^{2} y_{t-j}+\varepsilon_{t}
$$

A comparison of power series coefficients provides the identities:

$$
\begin{aligned}
\Phi & =-I_{p}+C(I-\bar{A})^{-1} B \\
\Psi & =-I_{p}-C(I-\bar{A})^{-2} \bar{A} B \\
\Pi_{j} & =\left[C \bar{A}^{2}(I-\bar{A})^{-2}\right] \bar{A}^{j-1} B=D \bar{A}^{j-1} B, j=1,2, \ldots
\end{aligned}
$$

It follows that the coefficients $\Pi_{j}, j=1,2, \ldots$ form the impulse response of a rational transfer function of order smaller or equal to $n$. If $\bar{A}$ is nonsingular then the order equals $n$ and the system $(\bar{A}, B, D)$ is minimal. Furthermore, it follows that for arbitrary $\Phi$ and $\Psi$ the transfer function

$$
a(z)=(1-z)^{2} I_{p}-\Phi z-\Psi z(1-z)-(1-z)^{2} z D(I-z \bar{A})^{-1} B
$$

is a rational transfer function with the additional property that

$$
a(1)=-\Phi=-\alpha \beta^{\prime}, \quad \bar{\alpha}_{\perp}^{\prime} \dot{a}(1) \bar{\beta}_{\perp}=\bar{\alpha}_{\perp}^{\prime}(-\Phi+\Psi) \bar{\beta}_{\perp}=-\bar{\alpha}_{\perp}^{\prime} C(I-\bar{A})^{-2} B \bar{\beta}_{\perp}=\zeta \eta^{\prime} .
$$

Consequently $\Phi$ and $\Psi$ determine the integration properties of processes generated using $a(z)$.
Conversely whenever the constraints

$$
-I_{p}+C(I-\bar{A})^{-1} B=\alpha \beta^{\prime}, \quad-\bar{\alpha}_{\perp}^{\prime} C(I-\bar{A})^{-2} B \bar{\beta}_{\perp}=\zeta \eta^{\prime}
$$

hold the corresponding triple $(A, B, C)$ corresponds to an $\mathrm{I}(2)$ process (if the eigenvalues of $A$ are in the closed unit disc). Defining $C_{*}=\bar{\alpha}^{\prime} C, C_{\dagger}=\bar{\alpha}_{\perp}^{\prime} C$ we obtain

$$
\begin{equation*}
-\bar{\alpha}^{\prime}+C_{*}(I-\bar{A})^{-1} B=\beta^{\prime}, \quad-\bar{\alpha}_{\perp}^{\prime}+C_{+}(I-\bar{A})^{-1} B=0, \quad-C_{+}(I-\bar{A})^{-2} B \bar{\beta}_{\perp}=\zeta \eta^{\prime} \tag{22}
\end{equation*}
$$

The third equation does not have a solution for fixed $B \bar{\beta}_{\perp}, \zeta, \eta$, if the row space of $B \bar{\beta}_{\perp}$ does not contain the space spanned by the rows of $\eta^{\prime}$. In this case row-wise projection of $\eta^{\prime}$ onto the space spanned by the rows of $B \bar{\beta}_{\perp}$ allows for (not necessarily unique) solutions in $C_{+}$. In the limit no projection is needed. Consequently for large enough $T$ the projected matrix will have full row rank. The second equation then determines $\bar{\alpha}_{\perp}$ which in turn determines $\bar{\alpha}$ up to the choice of the basis such that $\bar{\alpha}^{\prime}=T_{C} \overline{\alpha_{o}}{ }^{\prime}$ for some full row rank matrix $\overline{\alpha_{o}}{ }^{\prime} \in \mathbb{R}^{r \times p}, \overline{\alpha_{o}} \bar{\alpha}_{\perp}=0$. The first equation then can be rewritten as

$$
\left[T_{C}, C_{*}\right] \underbrace{\left[\begin{array}{c}
-\overline{\alpha_{o}^{\prime}} \\
(I-\bar{A})^{-1} B
\end{array}\right]}_{R_{1}}=\beta^{\prime} .
$$

The second equation shows that the row space of $(I-\bar{A})^{-1} B$ contains the row space of $\bar{\alpha}_{\perp}^{\prime}$. Thus the matrix $R_{1}$ has full row rank. It follows that this equation has solutions.

Having obtained a solution for $C_{*}, C_{\dagger}, \bar{\alpha}, \bar{\alpha}_{\perp}$ then $C$ is obtained from

$$
C=\left[\begin{array}{ll}
\alpha & \alpha_{\perp}
\end{array}\right]\left[\begin{array}{l}
C_{*} \\
C_{+}
\end{array}\right] .
$$

A unique solution then can be obtained from adding the restrictions $\Pi_{j}=C(I-\bar{A})^{-2} \bar{A}^{j+1} B, j=$ $1,2, \ldots, 2 n$ which for the estimates are to be solved in a least squares sense among all solutions to equations (22).

It then follows that for the true matrices $\Phi, \Psi, \Pi_{j}$ the only solution for given $\bar{A}, B$ consists in the corresponding true $C$. These facts therefore can be used in order to develop an initial guess for subsequent pseudo likelihood maximization using the parameterization of $\mathrm{I}(2)$ processes in state space representation: Given the integer valued parameters $n, c$ and $d$ :

1. Obtain a long VAR approximation $\hat{\Phi}, \hat{\Psi}^{\prime}, \hat{\Pi}_{j}, j=1,2, \ldots$, including $\hat{\Phi}=\hat{\alpha} \hat{\beta}^{\prime}$ and $\hat{\zeta} \hat{\eta}^{\prime}=\widehat{\alpha}_{\perp}^{\prime} \hat{\Psi} \widehat{\bar{\beta}_{\perp}}$ using the 2SI2 approach.
2. Choose the integer $f \geq n$. Use the algorithm described in Appendix F to obtain estimates $(\hat{A}, \hat{B}, \hat{D})$ realizing the impulse response $\bigcap_{j}, j=1, \ldots, 2 f$ from the Hankel matrix with $f$ block columns and $f$ block rows.
3. Project rows of $\hat{\eta}^{\prime}$ onto the space spanned by the rows of $\hat{B} \widehat{\bar{\beta}} \perp$ to obtain $\tilde{\eta}^{\prime}$.
4. Obtain a unique solution $\hat{C}$ solving (22) such that the matrices $\tilde{\Pi}_{j}=\hat{C}(I-\hat{A})^{-2} \hat{A}^{j+1} \hat{B}, j=$ $1,2, \ldots, 2 n$ have minimal Euclidean distance to $\hat{\Pi}_{j}, j=1,2, \ldots, 2 n$.
5. Transform the corresponding system $(\hat{A}+\hat{B} \hat{C}, \hat{B}, \hat{C})$ to the canonical form of Bauer and Wagner (2012) to obtain the estimate ( $\tilde{A}, \tilde{B}, \tilde{C})$.

The algorithm obtains a minimal state space system of order $n$ in the canonical form for $\mathrm{I}(2)$ processes given in Bauer and Wagner (2012) and hence can be used as an initial guess for subsequent pseudo-likelihood optimization in the set $M_{n}(r, s)$ of all order $n$ rational transfer functions corresponding to $\mathrm{I}(2)$ processes with state space unit root structure $((0,(c, c+d)))$.

Theorem 4 (Consistent initial guess). Let $\left(y_{t}\right)_{t \in \mathbb{Z}}$ denote a process generated using the system $\left(A_{0}, B_{0}, C_{0}\right)$ according to Assumption 3 and let the system ( $\tilde{A}, \tilde{B}, \tilde{C})$ be estimated based on the long VAR approximation with lag order chosen according to Assumption 2. Then $(\tilde{A}, \tilde{B}, \tilde{C})$ is a weakly consistent estimator of the data generating system $\left(A_{0}, B_{0}, C_{0}\right)$ in the sense that $\tilde{C} \tilde{A} \tilde{B} \xrightarrow{p} C_{0} A_{0}^{j} B_{0}, j=0,1, \ldots$ and hence the corresponding transfer functions converge in pointwise topology.

The proof of this theorem can be found in Appendix E.

## 6. Conclusions

In this paper the theory on long VAR approximation of general linear dynamical processes is extended to the case of $\mathrm{I}(2)$ processes. We find that we need slightly narrower upper and lower bounds in the approximations. The tighter bounds are not needed for all results and appear not very restrictive for applications.

The main results are completely analogous to the $\mathrm{I}(1)$ case: The asymptotics in many respects is identical to the finite order VAR case. Asymptotic distributions for the coefficients to non- stationary variables are the same as in the finite order VAR case. This holds true both for unrestricted OLS
estimates as well as the 2SI2 approach in the Johansen framework. Tests on cointegrating ranks show identical asymptotic distributions under the null as in the finite order VAR case and hence do not require other tables. In this respect the main conclusion is that the usual procedure of estimating the lag order in the first step and then applying the Johansen procedure for estimated lag order is justified also for processes generated from a $\operatorname{VAR}(\infty)$ that is approximated with a choice of the lag order lying within the prescribed bounds.

Additionally in the VARMA case the long VAR approximation can be used in order to derive consistent initial guesses that can be used in subsequent pseudo likelihood estimation.

Thus the paper provides both a full extension of results that have been achieved in the $I(1)$ case as well as a useful starting point for subsequent VARMA modeling which might be preferable in situations which require a high VAR order or show a large number of variables to be modeled, a situation where VARMA models can be more parsimonious than VAR models.

Author Contributions: The two authors of the paper have contributed equally, via joint efforts, regarding both ideas, research, and writing. Conceptualization, Y.L. and D.B.; methodology, Y.L. and D.B.; software, not applicable; validation, not applicable; formal analysis, Y.L. and D.B.; investigation, Y.L. and D.B.; resources, not applicable; writing-original draft preparation, Y.L. and D.B.; writing-review and editing, Y.L. and D.B.; visualization, not applicable; supervision, not applicable; project administration, D.B.; funding acquisition, D.B. Both authors have read and agreed to the published version of the manuscript.
Funding: This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation -Projektnummer 276051388) which is gratefully acknowledged. We acknowledge support for the publication costs by the Deutsche Forschungsgemeinschaft and the Open Access Publication Fund of Bielefeld University.
Acknowledgments: The reviewers and in particular the two guest editors provided significant comments that helped in improving the paper, which is gratefully acknowledged.
Conflicts of Interest: The authors declare no conflict of interest.

## Appendix A. Preliminaries

The theory in this paper follows closely the arguments in Lewis and Reinsel (1985) and its extension to the $I(1)$ case in Saikkonen and Lütkepohl (1996). To this end consider the finite order VECM approximation:

$$
\begin{equation*}
\Delta^{2} y_{t}=\Phi y_{t-1}+\Psi \Delta y_{t-1}+\sum_{j=1}^{h} \Pi_{j} \Delta^{2} y_{t-j}+e_{t} \tag{A1}
\end{equation*}
$$

The properties of the various estimators heavily use the following rewriting of the approximation using the triangular representation of $y_{t}$ :

$$
\begin{align*}
\Delta^{2} y_{t} & =\left[\Phi_{1}, \Phi_{2}, \Phi_{3}\right]\left[\begin{array}{c}
A \Delta y_{3, t-1}+u_{1, t-1} \\
y_{2, t-1} \\
y_{3, t-1}
\end{array}\right]+\left[\Psi_{1}, \Psi_{2}, \Psi_{3}\right]\left[\begin{array}{c}
A u_{3, t-1}+\Delta u_{1, t-1} \\
u_{2, t-1} \\
\Delta y_{3, t-1}
\end{array}\right] \\
& +\sum_{j=1}^{h}\left[\Pi_{j, 1}, \Pi_{j, 2}, \Pi_{j, 3}\right]\left[\begin{array}{c}
A \Delta u_{3, t-j}+\Delta^{2} u_{1, t-j} \\
\Delta u_{2, t-j} \\
u_{3, t-j}
\end{array}\right]+e_{t}  \tag{A2}\\
& =\Phi_{2} y_{2, t-1}+\Phi_{3} y_{3, t-1}+\Theta \Delta y_{3, t-1}+\sum_{j=1}^{h} \Xi_{j} u_{t-j}+\left[\Xi_{h+1,1}, \Xi_{h+1,2}\right]\left[\begin{array}{l}
u_{1, t-h-1} \\
u_{2, t-h-1}
\end{array}\right]+\Xi_{h+2,1} \tilde{u}_{1, t-h-2}+e_{t},
\end{align*}
$$

where $\tilde{u}_{1, t-h-2}:=u_{1, t-h-2}-A u_{3, t-h-1}$ and $\Phi_{2}=\Phi_{3}=0, \Theta=\Phi_{1} A+\Psi_{3}=0$, and

$$
\begin{array}{ll}
\Xi_{1} & =\left[\Phi_{1}+\Psi_{1}+\Pi_{1,1}, \Psi_{2}+\Pi_{1,2},\left(\Psi_{1}+\Pi_{1,1}\right) A+\Pi_{1,3}\right] \\
\Xi_{2} & =\left[-\Psi_{1}+\Pi_{2,1}-2 \Pi_{1,1}, \Pi_{2,2}-\Pi_{1,2},\left(\Pi_{2,1}-\Pi_{1,1}\right) A+\Pi_{2,3}\right] \\
\Xi_{j} & =\left[\Pi_{j, 1}-2 \Pi_{j-1,1}+\Pi_{j-2,1}, \Pi_{j, 2}-\Pi_{j-1,2},\left(\Pi_{j, 1}-\Pi_{j-1,1}\right) A+\Pi_{j, 3}\right], \quad j=3, \ldots, h, \\
\Xi_{h+1,1} & =-2 \Pi_{h, 1}+\Pi_{h-1,1}, \quad \Xi_{h+1,2}=-\Pi_{h, 2}, \quad \Xi_{h+2,1}=\Pi_{h, 1} .
\end{array}
$$

Furthermore, we can see that $\sum_{j=1}^{h+2} \Xi_{j, 1}=\Phi_{1}, \sum_{j=1}^{h+1} \Xi_{j, 2}=\Psi_{2}$, and $\sum_{j=1}^{h} \Xi_{j, 3}=\Psi_{1} A+\sum_{j=1}^{h} \Pi_{j, 3}$. Finally $\Psi_{1}=-\sum_{j=2}^{h+2}(j-1) \Xi_{j, 1}$.

Note that in the reparametrization (A2), the $\mathrm{I}(1)$ components, $y_{c, t}:=\left(y_{2, t}^{\prime}, \Delta y_{3, t}^{\prime}\right)^{\prime}$, as well as the $\mathrm{I}(2)$ components, $y_{3, t-1}$, are isolated from the stationary ones, $u_{t-j}$, and have coefficients equal to zero, which facilitates the derivation of the asymptotic properties.

In the reparameterized setting define ${ }^{3} \Xi:=\left[\Xi_{1}, \ldots, \Xi_{h}, \Xi_{h+1,1}, \Xi_{h+1,2}, \Xi_{h+2,1}\right], \quad p \times\left(p h+2 p_{1}+p_{2}\right)$, $U_{t}:=\left[u_{t-1}^{\prime}, \ldots, u_{t-h}^{\prime}, u_{1, t-h-1}^{\prime}, u_{2, t-h-1}^{\prime}, \tilde{u}_{1, t-h-2}^{\prime}\right]^{\prime}, \quad\left(p h+2 p_{1}+p_{2}\right) \times 1$,
$\underline{\Lambda}:=\left[\underline{\underline{\Xi}}, \Phi_{2}, \Theta, \Phi_{3}\right]=[\underline{\Xi}, 0], \quad p \times p(h+2)$,
$W_{t}:=\left[U_{t}^{\prime}, y_{c, t-1}^{\prime}, y_{3, t-1}^{\prime}\right]^{\prime}, \quad p(h+2) \times 1$.
we have

$$
\begin{equation*}
\Delta^{2} y_{t}=\underline{\Lambda} W_{t}+e_{t} \tag{A3}
\end{equation*}
$$

and correspondingly,

$$
\Delta^{2} y_{t}=\hat{\Lambda} W_{t}+\tilde{e}_{t}
$$

where

$$
\hat{\Lambda}=\left[\hat{\Xi}, \hat{\Phi}_{2}, \hat{\Theta}, \hat{\Phi}_{3}\right]=\left\langle\Delta^{2} y_{t}, W_{t}\right\rangle\left\langle W_{t}, W_{t}\right\rangle^{-1}
$$

is the OLS estimator of $\underline{\Lambda}$. Here $\left\langle X_{t}, Z_{t}\right\rangle:=\sum_{t=h+3}^{T} X_{t} Z_{t}^{\prime}$.
Note that $W_{t}$ and the regressors in (A1) are in one-one correspondence. In the original Equation (A1) beside the nonstationary regressors $y_{c, t-1}$ and $y_{3, t-1}$ the regressor vector

$$
\tilde{X}_{t}=\left[y_{1, t-1}^{\prime}, \Delta y_{1, t-1}^{\prime}, u_{2, t-1}^{\prime}, \Delta^{2} y_{t-1}^{\prime}, \ldots, \Delta^{2} y_{t-h}^{\prime}\right]^{\prime} \in \mathbb{R}^{2 p_{1}+p_{2}+p h}
$$

occurs which cointegrates with $\Delta y_{3, t-1}$ such that

$$
\begin{equation*}
X_{t}=\tilde{X}_{t}-\left[A^{\prime}, 0\right]^{\prime} \Delta y_{3, t-1}=T_{h} U_{t} \tag{A4}
\end{equation*}
$$

is stationary. Here the nonsingular matrix $T_{h} \in \mathbb{R}^{\left(p h+2 p_{1}+p_{2}\right) \times\left(p h+2 p_{1}+p_{2}\right)}$ is defined as:


Let $\underline{\Pi}:=\left[\Phi_{1}, \Psi_{1}, \Psi_{2},: \Pi_{1}: \dot{\Pi}_{2}: \ldots: \Pi_{h}\right]$, so that we have'

$$
\begin{equation*}
\underline{\Xi}=\underline{\Pi} T_{h} . \tag{A5}
\end{equation*}
$$

It can be verified that $T_{h}$ is invertible. The asymptotic properties of $\hat{\Lambda}-\underline{\Lambda}$ are clarified in the next lemma:

Lemma A1. Under the assumptions of Theorem 1 using $N=T-h-2$ as the effective sample size

[^27]\[

$$
\begin{aligned}
N^{\frac{1}{2}}(\hat{\Xi}-\underline{\Xi}) & =N^{\frac{1}{2}}\left\langle\varepsilon_{t}, U_{t}\right\rangle\left(\mathbb{E} U_{t} U_{t}^{\prime}\right)^{-1}+o_{P}\left(h^{\frac{1}{2}}\right), \\
{\left[N \hat{\Phi}_{2}, N \hat{\Theta}, N^{2} \hat{\Phi}_{3}\right] } & \Rightarrow g(1)\left[\begin{array}{ll}
\int_{0}^{1} d B B_{c}^{\prime} & \int_{0}^{1} d B H_{3}^{\prime}
\end{array}\right]\left[\begin{array}{cc}
\int_{0}^{1} B_{c} B_{c}^{\prime} & \int_{0}^{1} B_{c} H_{3}^{\prime} \\
\int_{0}^{1} H_{3} B_{c}^{\prime} & \int_{0}^{1} H_{3} H_{3}^{\prime}
\end{array}\right]^{-1}
\end{aligned}
$$
\]

where $H_{3}(u)=\int_{0}^{u} B_{3}(s) d s$.
Proof. The proof essentially shows that the coefficients corresponding to the stationary regressors and the ones corresponding to the integrated regressors asymptotically can be dealt with separately. Let $D_{T}:=\operatorname{diag}\left[N^{-\frac{1}{2}} I_{p h+2 p_{1}+p_{2}}, N^{-1} I_{p_{2}+p_{3}}, N^{-2} I_{p_{3}}\right]$. Note that $N^{\frac{1}{2}}(\hat{\Xi}-\underline{\Xi}), N\left[\hat{\Phi}_{2}, \hat{\Theta}\right]$, and $N^{2} \hat{\Phi}_{3}$ are the 1st, 2nd and 3rd column blocks of $(\hat{\Lambda}-\underline{\Lambda}) D_{T}^{-1}$, respectively. Moreover, we have

$$
(\hat{\Lambda}-\underline{\Lambda}) D_{T}^{-1}=\left\langle e_{t}, W_{t}\right\rangle D_{T}\left(D_{T}\left\langle W_{t}, W_{t}\right\rangle D_{T}\right)^{-1}
$$

Let $\hat{R}:=D_{T}\left\langle W_{t}, W_{t}\right\rangle D_{T}$, and define $R:=\operatorname{diag}\left[\Gamma_{u}, R_{2}\right]$, where $\Gamma_{u}=\mathbb{E}\left[U_{t} U_{t}^{\prime}\right]$, and

$$
R_{2}:=\left[\begin{array}{ll}
N^{-2}\left\langle y_{c, t-1}, y_{c, t-1}\right\rangle & N^{-3}\left\langle y_{c, t-1}, y_{3, t-1}\right\rangle \\
N^{-3}\left\langle y_{3, t-1}, y_{c, t-1}\right\rangle & N^{-4}\left\langle y_{3, t-1}, y_{3, t-1}\right\rangle
\end{array}\right] .
$$

Note that each block of the matrix $R_{2}$ is of order $O_{p}(1)$, and moreover, both $R_{2}$ and its limit are almost surely invertible, as there is no cointegration between $y_{c, t-1}$ and $y_{3, t-1}$ (see Lemma 3.1.1 in Chan and Wei (1988), and Sims et al. (1990)). Note that

$$
(\hat{\Lambda}-\underline{\Lambda}) D_{T}^{-1}-\left\langle\varepsilon_{t}, W_{t}\right\rangle D_{T} R^{-1}=\underbrace{\left\langle e_{1 t}, W_{t}\right\rangle D_{T} R^{-1}}_{=: E_{1}}+\underbrace{\left\langle e_{1 t}, W_{t}\right\rangle D_{T}\left(\hat{R}^{-1}-R^{-1}\right)}_{=: E_{2}}+\underbrace{\left\langle\varepsilon_{t}, W_{t}\right\rangle D_{T}\left(\hat{R}^{-1}-R^{-1}\right)}_{=: E_{3}} .
$$

Here $\left\langle\varepsilon_{t}, W_{t}\right\rangle D_{T} R^{-1}$ has the limits stated in the lemma since:

$$
\begin{aligned}
& N^{-1}\left\langle\varepsilon_{t}, y_{c, t-1}\right\rangle \Rightarrow g(1) \int_{0}^{1} d B B_{c}^{\prime}, \quad N^{-2}\left\langle\varepsilon_{t}, y_{3, t-1}\right\rangle \Rightarrow g(1) \int_{0}^{1} d B H_{3}^{\prime}, \\
& {\left[\begin{array}{ll}
N^{-2}\left\langle y_{c, t-1}, y_{c, t-1}\right\rangle & N^{-3}\left\langle y_{c, t-1}, y_{3, t-1}\right\rangle \\
N^{-3}\left\langle y_{3, t-1}, y_{c, t-1}\right\rangle & N^{-4}\left\langle y_{3, t-1}, y_{3, t-1}\right\rangle
\end{array}\right] \Rightarrow\left[\begin{array}{ll}
\int_{0}^{1} B_{c} B_{c}^{\prime} & \int_{0}^{1} B_{c} H_{3}^{\prime} \\
\int_{0}^{1} H_{3} B_{c}^{\prime} & \int_{0}^{1} H_{3} H_{3}^{\prime}
\end{array}\right] .}
\end{aligned}
$$

The lemma therefore holds, if $E_{1}=\left[o_{P}\left(h^{1 / 2}\right), o_{P}(1), o_{P}(1)\right], E_{2}=o_{P}(1), E_{3}=o_{P}(1)$ can be shown (where the blocks in $E_{1}$ correspond to the partitioning of $W_{t}$ into stationary, $\mathrm{I}(1)$ and $\mathrm{I}(2)$ components). For this it is sufficient to show:
(I) $\left\|\hat{R}^{-1}-R^{-1}\right\|_{1}=O_{P}\left(h / N^{\frac{1}{2}}\right)$
(II) $\left\|\left\langle e_{1 t}, W_{t}\right\rangle D_{T}\right\|=o_{P}\left(h^{1 / 2}\right)$ where $N^{-1}\left\langle e_{1 t}, y_{c, t-1}\right\rangle=o_{P}(1)$ and $N^{-2}\left\langle e_{1 t}, y_{3, t-1}\right\rangle=o_{P}(1)$
(III) $\left\|\left\langle\varepsilon_{t}, W_{t}\right\rangle D_{T}\right\|=O_{P}\left(h^{1 / 2}\right)$.

Here $\|.\|_{1}$ denotes the spectral norm of a matrix while $\|$.$\| denotes the Frobenius norm.$
(I) To see $\left\|\hat{R}^{-1}-R^{-1}\right\|_{1}=O_{p}\left(h / N^{\frac{1}{2}}\right)$, according to Lewis and Reinsel (1985), it is sufficient to show $\|\hat{R}-R\|_{1}=O_{p}\left(h / N^{\frac{1}{2}}\right),\left\|R^{-1}\right\|_{1}=O_{p}(1)$. Note that
$\hat{R}-R=\left[\begin{array}{ccc}N^{-1}\left\langle U_{t}, U_{t}\right\rangle-\Gamma_{u} & N^{-\frac{3}{2}}\left\langle U_{t}, y_{c, t-1}\right\rangle & N^{-\frac{5}{2}}\left\langle U_{t}, y_{3, t-1}\right\rangle \\ N^{-\frac{3}{2}}\left\langle y_{c, t-1}, U_{t}\right\rangle & 0 & 0 \\ N^{-\frac{5}{2}}\left\langle y_{3, t-1}, U_{t}\right\rangle & 0 & 0\end{array}\right]=:\left[\begin{array}{ccc}\hat{Q} & \hat{P}_{12} & \hat{P}_{13} \\ \hat{P}_{21} & 0 & 0 \\ \hat{P}_{31} & 0 & 0\end{array}\right]$,
then we have $\mathbb{E}\|\hat{R}-R\|_{1}^{2} \leq \mathbb{E}\|\hat{R}-R\|^{2}=\mathbb{E}\|\hat{Q}\|^{2}+2\left(\mathbb{E}\left\|\hat{P}_{12}\right\|^{2}+\mathbb{E}\left\|\hat{P}_{13}\right\|^{2}\right)$.

Now let $U_{t}^{o}:=\left[u_{t-1}^{\prime}, \ldots, u_{t-h-2}^{\prime}\right]^{\prime}$, then there exists a transformation $T^{u}$ of full row rank, such that $U_{t}=T^{u} U_{t}^{o}$, where $T^{u}$ is a $\left(p h+2 p_{1}+p_{2}\right) \times p(h+2)$ matrix:

Then, we have $\hat{Q}=T^{u} \hat{Q}^{o} T^{u^{\prime}}$, where $\hat{Q}^{o}=\frac{1}{N}\left\langle U_{t}^{o}, U_{t}^{o}\right\rangle-\mathbb{E}\left[U_{t}^{o} U_{t}^{o^{\prime}}\right]$; moreover, $\hat{P}_{1 i}=T^{u} \hat{P}_{1 i}^{o}$ for $i=2,3$, where $\hat{P}_{12}^{o}=N^{-\frac{3}{2}}\left\langle U_{t}^{o}, y_{c, t-1}\right\rangle, \hat{P}_{13}^{o}=N^{-\frac{5}{2}}\left\langle U_{t}^{o}, y_{3, t-1}\right\rangle$. Since $\left\|T^{u}\right\|_{1}=O(1), \hat{Q}$ and $\hat{P}_{1 i}$ have the same rate of convergence as $\hat{Q}^{o}$ and $\hat{P}_{1 i}^{o}$, respectively. From Saikkonen (1991) Lemma A.2. we know $\mathbb{E}\left\|\hat{Q}^{o}\right\|^{2}=O\left(h^{2} / N\right)$ and $\mathbb{E}\left\|\hat{P}_{12}^{o}\right\|^{2}=O(h / N)$ by direct calculation.

For $\hat{P}_{13}^{o}$ note that

$$
\mathbb{E}\left\|y_{3, t-1}\right\|^{2}=\mathbb{E}\left\|\sum_{j=1}^{t-1} \sum_{i=1}^{j} u_{3, i}\right\|^{2}=\mathbb{E}\left\|\sum_{i=1}^{t-1} i u_{3, t-1-i}\right\|^{2}=O\left(t^{3}\right) .
$$

Then analogous calculation as for $\hat{P}_{12}^{o}$ show that $\mathbb{E}\left\|\hat{P}_{13}^{o}\right\|^{2}=O(h / N)$. Concluding we obtain $\mathbb{E}\|\hat{R}-R\|_{1}^{2}=O\left(h^{2} / N\right)$ such that $\|\hat{R}-R\|_{1}=O_{P}\left(h / N^{\frac{1}{2}}\right)$.

To show $\left\|R^{-1}\right\|_{1}=O_{P}(1)$ note that $R^{-1}=\operatorname{diag}\left\{\Gamma_{u}^{-1}, R_{2}^{-1}\right\}$ where $\left\|\Gamma_{u}^{-1}\right\|_{1}=O(1)$ (see Lewis and Reinsel (1985), p. 397) and $\left\|R_{2}^{-1}\right\|_{1}=O_{P}(1)$, since $R_{2}$ is a.s. invertible and converges in distribution to an almost surely nonsingular random matrix.
(II) With respect to $\left\|\left\langle e_{1 t}, W_{t}\right\rangle D_{T}\right\|=o_{P}\left(h^{1 / 2}\right)$ note that

$$
\left\|\left\langle e_{1 t}, W_{t}\right\rangle D_{T}\right\| \leq\left\|N^{-\frac{1}{2}}\left\langle e_{1 t}, U_{t}\right\rangle\right\|+\left\|N^{-1}\left\langle e_{1 t}, y_{c, t-1}\right\rangle\right\|+\left\|N^{-2}\left\langle e_{1 t}, y_{3, t-1}\right\rangle\right\| .
$$

From Saikkonen (1991) Lemma A. 5 we have $\left\|N^{-\frac{1}{2}}\left\langle e_{1 t}, U_{t}\right\rangle\right\|=o_{P}\left(h^{\frac{1}{2}}\right)$, and $\left\|N^{-1}\left\langle e_{1 t}, y_{c, t-1}\right\rangle\right\|=$ $o_{P}(1)$. Then $\mathbb{E}\left\|y_{3, t-1}\right\|^{2}=O\left(t^{3}\right)$ and $\mathbb{E}\left\|e_{1 t}\right\|^{2}=o\left(N^{-1}\right)$ imply

$$
\mathbb{E}\left\|N^{-2}\left\langle e_{1 t}, y_{3, t-1}\right\rangle\right\| \leq N^{-2} \sum_{t=h+3}^{T}\left(\mathbb{E}\left\|e_{1 t}\right\|^{2} \mathbb{E}\left\|y_{3, t-1}\right\|^{2}\right)^{\frac{1}{2}}=o\left(N^{-2} N N^{-1 / 2} N^{3 / 2}\right)=o(1)
$$

(III) To show $\left\|\left\langle\varepsilon_{t}, W_{t}\right\rangle D_{T}\right\|=O_{P}\left(h^{1 / 2}\right)$ note that $N^{-\frac{1}{2}}\left\langle\varepsilon_{t}, U_{t}\right\rangle=O_{P}\left(h^{1 / 2}\right), N^{-1}\left\langle\varepsilon_{t}, y_{c, t-1}\right\rangle=O_{P}(1)$ according to (A.7) of Saikkonen (1992). Moreover $N^{-2} \sum_{t=h+3}^{T} \varepsilon_{t} y_{3, t-1}^{\prime} \Rightarrow g(1) \int_{0}^{1} d B H_{3}^{\prime}$ implies $N^{-2}\left\langle\varepsilon_{t}, y_{3, t-1}\right\rangle=O_{P}(1)$.

Note that for the lemma to hold we only need $h^{3} / N \rightarrow 0$ and $N^{1 / 2} \sum_{j=h+1}^{\infty}\left\|\Pi_{j}\right\|=o(1)$.

## Appendix B. Proof of Theorem 1

Appendix B.1. (A) Consistency
(i) Lemma A1 implies $\hat{\Phi}_{2} \rightarrow 0=\Phi_{2}, \hat{\Phi}_{3} \rightarrow 0=\Phi_{3}$. Furthermore, the reparameterization implies $\Phi_{1}=\sum_{j=1}^{h+2} \Xi_{j 1}$ and thus $\hat{\Phi}_{1}=\sum_{j=1}^{h+2} \hat{\Xi}_{j, 1}$ leading to

$$
\begin{aligned}
\left\|\hat{\Phi}_{1}-\underline{\Phi}_{1}\right\| & \leq\left\|\sum_{j=1}^{h+2} \hat{\Xi}_{j, 1}-\sum_{j=1}^{h+2} \Xi_{j, 1}\right\| \\
& \leq \sum_{j=1}^{h+2}\left\|\hat{\Xi}_{j, 1}-\underline{\Xi}_{j, 1}\right\| \leq\|\hat{\Xi}-\underline{\Xi}\|=O_{P}\left(h^{3 / 2} / N^{1 / 2}\right)
\end{aligned}
$$

where the last inequality holds due to $\left\langle\varepsilon_{t}, u_{t-j}\right\rangle=O_{P}\left(N^{1 / 2}\right)$ in combination with Lemma A1.
(ii) Note that

$$
\hat{\Sigma}_{\epsilon}=N^{-1}\left\langle\Delta^{2} y_{t}-\hat{\Lambda} W_{t}, \Delta^{2} y_{t}-\hat{\Lambda} W_{t}\right\rangle=N^{-1}\left\langle e_{t}+(\underline{\Lambda}-\hat{\Lambda}) W_{t}, e_{t}+(\underline{\Lambda}-\hat{\Lambda}) W_{t}\right\rangle
$$

Now

$$
\left\langle(\underline{\Lambda}-\hat{\Lambda}) W_{t},(\underline{\Lambda}-\hat{\Lambda}) W_{t}\right\rangle=(\underline{\Lambda}-\hat{\Lambda}) D_{T}^{-1} D_{T}\left\langle W_{t}, W_{t}\right\rangle D_{T} D_{T}^{-1}(\underline{\Lambda}-\hat{\Lambda})^{\prime}
$$

where $\hat{R}=D_{T}\left\langle W_{t}, W_{t}\right\rangle D_{T}$ such that $\|\hat{R}\|_{1}=O_{P}(1)$ and $\left\|(\underline{\Lambda}-\hat{\Lambda}) D_{T}^{-1}\right\|=O_{P}\left(h^{1 / 2}\right)$. Consequently

$$
N^{-1}\left\langle(\underline{\Lambda}-\hat{\Lambda}) W_{t},(\underline{\Lambda}-\hat{\Lambda}) W_{t}\right\rangle=O_{P}(h / N) \rightarrow 0
$$

Next, from the definition of $e_{t}$, we can show that

$$
N^{-1}\left\langle\varepsilon_{t}+e_{1 t}, \varepsilon_{t}+e_{1 t}\right\rangle=N^{-1}\left\langle\varepsilon_{t}, \varepsilon_{t}\right\rangle+o_{P}(1)=\Sigma_{\epsilon}+o_{P}(1)
$$

where the last equality follows the law of large numbers and the first equality is implied by the fact that $\left\|e_{1 t}\right\|^{2}=o_{P}\left(T^{-1}\right)$ and $\left\|\varepsilon_{t}\right\|^{2}=O_{P}(1)$.
(iii) From (i) and (ii), $\hat{\Omega}_{1 . c}=\left(\hat{\Phi}_{1}^{\prime} \hat{\Sigma}_{\epsilon}^{-1} \hat{\Phi}_{1}\right)^{-1}=\left(\Phi_{1}^{\prime} \Sigma_{\epsilon}^{-1} \Phi_{1}\right)^{-1}+o_{P}(1)=\Omega_{1 . c}+o_{P}(1)$ directly follows.
(iv) With respect to $\hat{\Psi}$ recall that

$$
\Psi_{1}=-\sum_{j=2}^{h+2}(j-1) \Xi_{j, 1}, \quad \Psi_{2}=\sum_{j=1}^{h+1} \Xi_{j, 2} .
$$

Then Lemma A1 shows that each entry of $\hat{E}-\underline{E}$ is of order $O_{P}\left(h^{1 / 2} / N^{1 / 2}\right)$. Then

$$
\left\|\hat{\Psi}_{1}-\Psi_{1}\right\| \leq \sum_{j=2}^{h+2}(j-1)\left\|\hat{\Xi}_{j, 1}-\Xi_{j, 1}\right\|=O_{P}\left(\sum_{j=2}^{h+2}(j-1) h^{1 / 2} / N^{1 / 2}\right)=O_{P}\left(h^{5 / 2} / N^{1 / 2}\right)
$$

which converges to zero for $h^{5} / T \rightarrow 0$. Similarly $\hat{\Psi}_{2}-\Psi_{2}=O_{P}\left(h^{3 / 2} / N^{1 / 2}\right)$.
For $\hat{\Psi}_{3}$ note that $\Theta=\Phi_{1} A+\Psi_{3}$. Thus $\hat{\Psi}_{3}=\hat{\Theta}-\hat{\Phi}_{1} A$ such that $\hat{\Psi}_{3} \rightarrow \Psi_{3}$ from (i) and Lemma A1.
(v) is contained in Lemma A1.
(vi) From (6), and the definition $\hat{\Omega}_{1 . c}=\left(\hat{\Phi}_{1}^{\prime} \hat{\Sigma}_{\epsilon}^{-1} \hat{\Phi}_{1}\right)^{-1}$, we have

$$
\begin{aligned}
\hat{A}-A & =-\left(\hat{\Phi}_{1}^{\prime} \hat{\Sigma}_{\epsilon}^{-1} \hat{\Phi}_{1}\right)^{-1} \hat{\Phi}_{1}^{\prime} \hat{\Sigma}_{\epsilon}^{-1} \hat{\Psi}_{3}-A \\
& =-\hat{\Omega}_{1 . c} \hat{\Phi}_{1}^{\prime} \hat{\Sigma}_{\epsilon}^{-1} \hat{\Psi}_{3}-\hat{\Omega}_{1 . c} \hat{\Omega}_{1 . c}^{-1} A=-\hat{\Omega}_{1 . c} \hat{\Phi}_{1}^{\prime} \hat{\Sigma}_{\epsilon}^{-1} \hat{\Psi}_{3}-\hat{\Omega}_{1 . c} \hat{\Phi}_{1}^{\prime} \hat{\Sigma}_{\epsilon}^{-1} \hat{\Phi}_{1} A \\
& =-\hat{\Omega}_{1 . c} \hat{\Phi}_{1}^{\prime} \hat{\Sigma}_{\epsilon}^{-1}\left(\hat{\Psi}_{3}+\hat{\Phi}_{1} A\right)=-\hat{\Omega}_{1 . c} \hat{\Phi}_{1}^{\prime} \hat{\Sigma}_{\epsilon}^{-1} \hat{\Theta} .
\end{aligned}
$$

Then (i-iii, v) show the result.
Appendix B.2. (B) Asymptotic Distribution of Coefficients to Nonstationary Regressors
(i) The distribution of the coefficients due to the nonstationary components is contained in Lemma A1.
(ii) With respect to the cointegrating relation note that from the proof of Theorem 1 we have

$$
N(\hat{A}-A)=-N \hat{\Omega}_{1 . c} \hat{\Phi}_{1}^{\prime} \hat{\Sigma}_{\epsilon}^{-1} \hat{\Theta}=-\Omega_{1 . c} \Phi_{1}^{\prime} \Sigma_{\epsilon}^{-1} \cdot N \hat{\Theta}+o_{P}(1)
$$

Note that $N \hat{\Theta}=\left[N \hat{\Phi}_{2}, N \hat{\Theta}, N^{2} \hat{\Phi}_{3}\right] \eta$, where $\eta=\left[0_{p_{3} \times p_{2}}, I_{p_{3}}, 0_{p_{3} \times p_{3}}\right]^{\prime}$. Then by Lemma A1, we have
$N(\hat{A}-A) \Rightarrow-\Omega_{1 . c} \Phi_{1}^{\prime} \Sigma_{\epsilon}^{-1} \cdot g(1) \int_{0}^{1} d B F^{\prime}\left(\int_{0}^{1} F F^{\prime}\right)^{-1} \eta=-\Omega_{1 . c} \Phi_{1}^{\prime} \Sigma_{\epsilon}^{-1} \cdot g(1) \int_{0}^{1} d B L^{\prime}\left(\int_{0}^{1} L L^{\prime}\right)^{-1}$.
Note that $\Phi_{1}=g(1) \alpha$, and by definition $\Omega=\left[\begin{array}{ll}\Omega_{11} & \Omega_{1 c} \\ \Omega_{c 1} & \Omega_{c c}\end{array}\right]=g(1)^{-1} \Sigma_{\epsilon} g(1)^{\prime-1}$, we have

$$
\begin{aligned}
-\Omega_{1 . c} \Phi_{1}^{\prime} \Sigma_{\epsilon}^{-1} g(1) B & =-\Omega_{1 . c} \alpha^{\prime} g(1)^{\prime} \Sigma_{\epsilon}^{-1} g(1) B=-\Omega_{1 . c}\left[I_{p_{1}} 0\right] \Omega^{-1} B \\
& =\Omega_{1 . c}\left[\left(\Omega^{-1}\right)_{11}\left(\Omega^{-1}\right)_{1 c}\right] B=\Omega_{1 . c}\left[\Omega_{1 . c}^{-1} \quad-\Omega_{1 . c}^{-1} \Omega_{1 c} \Omega_{c c}^{-1}\right] B \\
& =\left[\begin{array}{ll}
I_{p_{1}} & -\Omega_{1 c} \Omega_{c c}^{-1}
\end{array}\right]\left[\begin{array}{l}
B_{1} \\
B_{c}
\end{array}\right]=B_{1}-\Omega_{1 c} \Omega_{c c}^{-1} B_{c}=B_{1 . c} .
\end{aligned}
$$

Therefore, we have

$$
N(\hat{A}-A) \Rightarrow \int_{0}^{1} d B_{1 . c} L^{\prime}\left(\int_{0}^{1} L L^{\prime}\right)^{-1}
$$

Appendix B.3. (C) Asymptotic Distribution of Coefficients to Stationary Regressors
Since the regressor vector $U_{t}$ is stationary, the asymptotic distribution of $N^{1 / 2} L_{h}^{\prime} \operatorname{vec}(\hat{\Xi}-\underline{\Xi})$ follows from Lewis and Reinsel (1985) in combination with uniform boundedness of the maximal and the minimal eigenvalue of $\Gamma_{u}=\mathbb{E} U_{t} U_{t}^{\prime}$, see above. Analogously the result for the coefficients corresponding to the regressor vector $X_{t}$ are shown as $X_{t}=T_{h} U_{t}$ for nonsingular matrix $T_{h}$.

Appendix B.4. (D) Asymptotic Distribution of Wald Type Tests
For the Wald test in addition to (C) note that the variance $\Gamma_{E C M}$ is replaced by an estimate $\hat{\Gamma}_{E C M}$. For

$$
L_{h}^{\prime}\left(\Gamma_{E C M}^{-1} \otimes \Sigma_{\epsilon}\right) L_{h}-L_{h}^{\prime}\left(\hat{\Gamma}_{E C M}^{-1} \otimes \hat{\Sigma}_{\epsilon}\right) L_{h}
$$

note that $\hat{\Sigma}_{\epsilon}-\Sigma_{\epsilon}=o_{P}(1)$ due to (A) (ii). The regressor vectors $\tilde{X}_{t}$ and $X_{t}$ differ only in the first block where $y_{1, t-1}=u_{1, t-1}+A \Delta y_{3, t-1}$ replaces $u_{1, t-1}$. Regressing out $\Delta y_{3, t-1}$ eliminates this difference. Then $\left\|\hat{\Gamma}_{E C M}-\Gamma_{E C M}\right\|_{1}=O_{P}\left(h / N^{1 / 2}\right)$ according to (Saikkonen and Lütkepohl 1996, p. 835, 1. 3). There also invertibility of $\Gamma_{E C M}$ is shown. Using Lemma A. 2 of Saikkonen and Lütkepohl (1996) this implies $\left\|\hat{\Gamma}_{E C M}^{-1}-\Gamma_{E C M}^{-1}\right\|_{1}=O_{P}\left(h / N^{1 / 2}\right)$.

The rest then follows as the proof of Theorem 4 in Saikkonen and Lütkepohl (1996).

## Appendix C. Proof of Theorem 2

Consistency follows directly from Theorem 1 as the general representation can be transformed into a triangular representation using the matrix $\mathcal{B}=\left[\beta, \beta_{1}, \beta_{2}\right]$, see (4).

With respect to the asymptotic distribution following the proof of Theorem 1 there exists a nonsingular transformation matrix $S_{h}$ such that $W_{t}=S_{h} Z_{t, h}$. From $\left\|\hat{R}^{-1}-R^{-1}\right\|=O_{P}\left(h / N^{1 / 2}\right)$ it follows that

$$
\left(N^{-1}\left\langle W_{t}, W_{t}\right\rangle\right)^{-1}=\left[\begin{array}{cc}
\left(\Gamma_{u}\right)^{-1} & 0 \\
0 & 0
\end{array}\right]+o_{P}\left(h / N^{1 / 2}\right)
$$

Therefore it follows that the blocks corresponding to the nonstationary regressors do not contribute to the asymptotic distribution. Then standard arguments for the stationary part of the regressor vector can be used.

## Appendix D. Proofs for Theorem 3

The proof combines the ideas of Saikkonen and Luukkonen (1997) (in the following S\&L) with the asymptotics of 2SI2 of Paruolo (2000) (in the following P). In the proof we will work without restriction of generality with the triangular representation.

The key to the asymptotic properties of the estimators obtained from the 2SI2 algorithm lies in the results of P Lemma A. 4 and Lemma A. 5 in the appendix. These lemmas deal with the limits of various moment matrices of the form $N^{-a}\left\langle R_{i t}, R_{j t}\right\rangle$ corrected for the stationary components $\Delta^{2} y_{t-j}, j=1, \ldots, h-2$. The correction involves a regressor vector growing in dimension with sample size. This is dealt with in S\&L.

In this respect let $S_{t}=\left[\Delta^{2} y_{t-1}^{\prime}, \ldots, \Delta^{2} y_{t-h+2}^{\prime}\right]^{\prime}$ which according to (A4) is a linear function of $U_{t}$ such that $S_{t}=\mathcal{T}_{s} U_{t}$. The definition of $U_{t}$ implies $\hat{Q}=N^{-1}\left\langle U_{t}, U_{t}\right\rangle-\mathbb{E} U_{t} U_{t}^{\prime}=O_{P}\left(h / N^{1 / 2}\right)$. On p. 543 in P the matrices $\Sigma_{i j}, i, j \in\{Y, U, 0\}$ are defined as limits of second moment matrices. Here ${ }^{\prime} U^{\prime}$ refers to $\beta_{1}^{\prime} \Delta y_{t-1}=u_{2, t-1}$ in the triangular representation, ' $Y^{\prime}$ refers to $\beta^{\prime} y_{t-1}+\delta \beta_{2}^{\prime} \Delta y_{t-1}=$ $y_{1, t-1}-A \Delta y_{3, t-1}=u_{1, t-1}$ and ' $0^{\prime}$ refers to $\Delta^{2} y_{t}$. These are all stationary processes and linear functions of $u_{t}, u_{t-1}, u_{t-2}$. Additional to $S_{t}$ also $\beta^{\prime} \Delta y_{t-1}=\Delta u_{1, t-1}+A u_{3, t-1}$ is corrected for in the second stage.

The arguments on p. 114 and 115 of S\&L deal with terms of the form

$$
N^{-1}\left\langle u_{1, t-1}, u_{1, t-1}\right\rangle-N^{-1}\left\langle u_{1, t-1}, S_{t}\right\rangle\left\langle S_{t}, S_{t}\right\rangle^{-1}\left\langle S_{t}, u_{1, t-1}\right\rangle .
$$

Analogous arguments to $\mathrm{S} \& \mathrm{~L}(\mathrm{~A} .12)$ show that this equals (up to terms of order $o_{P}(1)$ )

$$
C_{11}=\mathbb{E} u_{1, t-1} u_{1, t-1}^{\prime}-\mathbb{E} u_{1, t-1} S_{t}^{\prime}\left(\mathbb{E} S_{t} S_{t}^{\prime}\right)^{-1} \mathbb{E} S_{t} u_{1, t-1}^{\prime}
$$

S\&L state that this is bounded from above and bounded away from zero. The second claim actually is wrong. If $\left(u_{1, t}\right)_{t \in \mathbb{Z}}$ is univariate white noise with unit variance then $C_{11}=\frac{1}{h}$ is achieved by predicting $u_{1, t-1}$ by

$$
\sum_{j=1}^{h} \frac{h-j}{h} \Delta u_{1, t-j}=u_{1, t-1}-\frac{1}{h} \sum_{j=1}^{h} u_{1, t-j}
$$

including integration of the regressors in the form of the summation. This does not change the remaining arguments in S\&L, it only implies that the separation of the eigenvalues corresponding to the stationary regressors and the ones corresponding to the non-stationary ones is weaker.

In the current case one can show that for

$$
N^{-1}\left\langle u_{1, t-1}, u_{1, t-1}\right\rangle-N^{-1}\left\langle u_{1, t-1}, S_{t}\right\rangle\left\langle S_{t}, S_{t}\right\rangle^{-1}\left\langle S_{t}, u_{1, t-1}\right\rangle
$$

where $S_{t}$ contains $\Delta u_{1, t-1}$ and $\Delta^{2} u_{1, t-j}, j=1, \ldots, h$ for the corresponding limit $C_{11}$ the lower bound $h C_{11} \geq c I$ holds for some $0<c$. The order of the lower bound is achieved by including a double integration of the regressors. For

$$
N^{-1}\left\langle\Delta u_{1, t-1}, \Delta u_{t, t-1}\right\rangle-N^{-1}\left\langle\Delta u_{1, t-1}, S_{t}\right\rangle\left\langle S_{t}, S_{t}\right\rangle^{-1}\left\langle S_{t}, \Delta u_{1, t-1}\right\rangle=C_{\Delta \Delta}+o_{p}(1)
$$

we have $h^{3} C_{\Delta \Delta} \geq c I$. Here the arguments from above can be applied to the process $\left(\Delta u_{t}\right)_{t \in \mathbb{Z}}$. For a differenced process the smallest eigenvalue of the matrix

$$
\mathbb{E} \delta U_{t} \delta U_{t}^{\prime}, \quad \delta U_{t}^{\prime}=\left[\Delta u_{t}^{\prime}, \Delta u_{t-1}^{\prime}, \ldots, \Delta u_{t-h}^{\prime}\right]
$$

is of order $h^{-2}$, compare Theorem 2 of Palma and Bondon (2003).

Since $N^{-1}\left\langle S_{t}, y_{c, t-1}\right\rangle=O_{P}\left(h^{1 / 2}\right)$ and $N^{-2}\left\langle S_{t}, y_{3, t-1}\right\rangle=O_{P}\left(h^{1 / 2}\right)$ it follows that

$$
\begin{aligned}
N^{-1}\left(\left\langle u_{1, t-1}, y_{c, t-1}\right\rangle-\left\langle u_{1, t-1}, S_{t}\right\rangle\left\langle S_{t}, S_{t}\right\rangle^{-1}\left\langle S_{t}, y_{c, t-1}\right\rangle\right) & =O_{P}\left(h^{1 / 2}\right) \\
N^{-2}\left(\left\langle y_{c, t-1}, y_{c, t-1}\right\rangle-\left\langle y_{c, t-1}, S_{t}\right\rangle\left\langle S_{t}, S_{t}\right\rangle^{-1}\left\langle S_{t}, y_{c, t-1}\right\rangle\right) & =N^{-2}\left\langle y_{c, t-1}, y_{c, t-1}\right\rangle+o_{P}\left((h / N)^{1 / 2}\right)
\end{aligned}
$$

as well as

$$
\begin{aligned}
& N^{-2}\left(\left\langle u_{1, t-1}, y_{3, t-1}\right\rangle-\left\langle u_{1, t-1}, S_{t}\right\rangle\left\langle S_{t}, S_{t}\right\rangle^{-1}\left\langle S_{t}, y_{3, t-1}\right\rangle\right)=O_{P}\left(h^{1 / 2}\right), \\
& N^{-3}\left(\left\langle y_{c, t-1}, y_{3, t-1}\right\rangle-\left\langle y_{c, t-1}, S_{t}\right\rangle\left\langle S_{t}, S_{t}\right\rangle^{-1}\left\langle S_{t}, y_{3, t-1}\right\rangle\right)=N^{-3}\left\langle y_{c, t-1}, y_{3, t-1}\right\rangle+o_{P}\left((h / N)^{1 / 2}\right), \\
& N^{-4}\left(\left\langle y_{3, t-1}, y_{3, t-1}\right\rangle-\left\langle y_{3, t-1}, S_{t}\right\rangle\left\langle S_{t}, S_{t}\right\rangle^{-1}\left\langle S_{t}, y_{3, t-1}\right\rangle\right)=N^{-4}\left\langle y_{3, t-1}, y_{3, t-1}\right\rangle+o_{P}\left((h / N)^{1 / 2}\right) .
\end{aligned}
$$

Therefore the limits of the moment matrices $M_{i j}$ are not affected by the correction using stationary terms even if $h \rightarrow \infty$ except for the terms involving the orders $O_{P}\left(h^{1 / 2}\right)$. For all stationary terms we find convergence to the corresponding limits denoted $\Sigma_{i j}$ in P .

The first step in the 2SI2 procedure then uses RRR in the equation

$$
\Delta^{2} y_{t}=\Psi \Delta y_{t-1}+\alpha \beta^{\prime} y_{t-1}+\underline{\Pi} S_{t}+e_{t} .
$$

Then $R_{0 t}$ denotes $\Delta^{2} y_{t}$ corrected for $S_{t}, R_{1, t}$ denotes $\Delta y_{t-1}$ corrected for $S_{t}$ and $R_{2, t}$ denotes $y_{t-1}$ corrected for $S_{t}$. Lemma A. 4 of P derives the limits of different directions of $M_{i j . k}$ defined as

$$
M_{i j . k}=M_{i j}-M_{i k} M_{k k}^{-1} M_{k j}, M_{i j}=N^{-1}\left\langle R_{i, t}, R_{j, t}\right\rangle
$$

where $i, j \in\{0,1,2, \varepsilon, \beta\}$. Here $R_{\varepsilon, t}$ equals $e_{t}$ correct for $S_{t}$ and $R_{\beta, t}=\beta^{\prime} R_{1, t}$. Further P uses the notation $A_{T}=\left[\bar{\beta}_{1}, T^{-1} \bar{\beta}_{2}\right]$ and $\bar{\beta}_{2, T}=\bar{\beta}_{2}$. Here and below we assume without restriction of generality that $\left[\beta, \beta_{1}, \beta_{2}\right]$ is an orthonormal matrix. Consequently $\bar{\beta}=\beta, \overline{\beta_{1}}=\beta_{1}, \bar{\beta}_{2}=\beta_{2}$. Then the results above imply all results of Lemma A.4. of P except that now $A_{T}^{\prime} M_{20.1}=O_{P}\left(h^{1 / 2}\right)$.

In particular we obtain the following limits:

$$
\begin{array}{lll}
A_{T}^{\prime} M_{2 \varepsilon .1} \xrightarrow{d} \int_{0}^{1} F_{+}(d W)^{\prime} \quad, & A_{T}^{\prime} M_{22.1} A_{T} \xrightarrow{d} \int_{0}^{1} F_{+} F_{+}^{\prime}, \\
\beta_{2}^{\prime} M_{1 \varepsilon . \beta} \xrightarrow{d} \int_{0}^{1} B_{3}(d W)^{\prime} \quad, & T^{-1} \beta_{2}^{\prime} M_{11 . \beta} \beta_{2} \xrightarrow{d} \int_{0}^{1} B_{3} B_{3 \prime}^{\prime} \\
\beta_{2}^{\prime} M_{1 \varepsilon . b} \xrightarrow{d} \int_{0}^{1} L(d W)^{\prime} \quad, & T^{-1} \beta_{2}^{\prime} M_{11 . b} \beta_{2} \xrightarrow{d} \int_{0}^{1} L L^{\prime} .
\end{array}
$$

Here $W=g(1) B$ denotes the Brownian motion corresponding to $\left(\varepsilon_{t}\right)_{t \in \mathbb{Z}}, F_{+}$denotes the Brownian motion corresponding to $R_{2 t}$ (equaling $y_{t-1}$ corrected for $S_{t}$ ) corrected for $R_{1 t}\left(\Delta y_{t-1}\right.$ whose only nonstationary component equals $\Delta y_{3, t-1}$ with corresponding Brownian motion $B_{3}$ ). Thus we obtain the following definitions (where $L$ is as in Theorem 1):

$$
\begin{aligned}
F_{a}(u) & =\left[\begin{array}{c}
B_{2}(u) \\
\int_{0}^{u} B_{3}(v) d v
\end{array}\right], \quad F_{+}(u)=F_{a}(u)-\int_{0}^{1} F_{a} B_{3}^{\prime}\left(\int_{0}^{1} B_{3} B_{3}^{\prime}\right)^{-1} B_{3}(u), \\
L(u) & =B_{3}(u)-\int_{0}^{1} B_{3} F_{a}^{\prime}\left(\int_{0}^{1} F_{a} F_{a}^{\prime}\right)^{-1} F_{a}(u) .
\end{aligned}
$$

The above arguments show that in the current setting $U_{t-1}=u_{2, t-1}$ and $Y_{t-1}=u_{1, t-1}$ are contained in the space spanned by $S_{t}$ for $h \rightarrow \infty$. Therefore $\Sigma_{i j}=0$ for $i, j \in\{U, Y\}$. The subscript 'b' refers to correcting for $\beta_{\perp}^{\prime} R_{2 t}$ used in the second stage of 2SI2.

Let $\tilde{\Sigma}_{Y Y}$ denote the limit of $h\left\langle Y_{t-1}, Y_{t-1}\right\rangle$ and analogously define $\tilde{\Sigma}_{Y U}, \tilde{\Sigma}_{U U}, \tilde{\Sigma}_{0 Y}$ and $\tilde{\Sigma}_{0 U}$. For the latter two note that $\tilde{\Sigma}_{0 Y}$ denotes the limit of

$$
h\left\langle\Delta^{2} y_{t}, Y_{t-1}\right\rangle=h \alpha\left\langle Y_{t-1}, Y_{t-1}\right\rangle+h\left\langle\zeta U_{t-1}, Y_{t-1}\right\rangle+h\left\langle\zeta_{2} \beta^{\prime} \Delta y_{t-1}, Y_{t-1}\right\rangle+h \underline{\Pi}\left\langle S_{t}, Y_{t-1}\right\rangle+h\left\langle e_{t}, Y_{t-1}\right\rangle
$$

corrected for $S_{t}$ and $\beta^{\prime} \Delta y_{t-1}$. Since $Y_{t-1}$ is stationary the last term is of order $O_{P}\left(\left(h^{3} / N\right)^{1 / 2}\right)=o_{P}(1)$. Therefore it follows that $\tilde{\Sigma}_{0 Y}=\alpha \tilde{\Sigma}_{Y Y}+\zeta \tilde{\Sigma}_{U Y}$. Then the results of Lemma A. 5 of P hold where in (A.11) and (A.14) $\Sigma_{i j}$ can be replaced by $\tilde{\Sigma}_{i j}$.

The asymptotic analysis below will heavily use the Johansen approach of investigating the solutions to eigenvalue problems in order to maximize the pseudo-likelihood corresponding to the reduced rank regression problem. In order to use the corresponding local analysis one has to first clarify consistency for the various estimators as well as rates of convergence.

The main tool in this respect is Theorem A. 1 of Johansen (1997) which establishes in the I(2) setting for the regression $y_{t}=\theta^{\prime} Z_{t}+\varepsilon_{t}\left(Z_{t}\right.$ being composed of stationary, $\mathrm{I}(1)$ and $\mathrm{I}(2)$ components) where $D_{T}\left\langle Z_{t}, Z_{t}\right\rangle D_{T}=O_{P}(1)$ and $D_{T}\left\langle Z_{t}, \varepsilon_{t}\right\rangle=o_{P}(1)$ that $D_{T}^{-1}(\hat{\theta}-\theta)=o_{P}(1)$ where $\hat{\theta}$ denotes the pseudo likelihood estimator over some closed parameter set $\Theta$.

It is straightforward to see that analogous results hold in the present setting when first concentrating out the stationary components: Consider $y_{t}=\theta_{1}^{\prime} z_{t}+\theta_{2}^{\prime} Z_{t}+e_{t}$. Then $\hat{\theta}_{2}\left(\theta_{1}\right)$ is obtained from the concentration step and the pseudo likelihood involves $\left\langle R_{t, y}-\theta_{1}^{\prime} R_{t, z}, R_{t, y}-\theta_{1}^{\prime} R_{t, z}\right\rangle$ where again the processes $R_{t, y}$ and $R_{t, z}$ denote the processes $y_{t}$ and $z_{t}$ with the corresponding stationary regressors $Z_{t}$ regressed out. These concentrated quantities now can be used in the proof of Theorem A. 1 of Johansen (1997) essentially without changes to show consistency for $\hat{\theta}_{1}$. Consistency of $\hat{\theta}_{2}\left(\hat{\theta}_{1}\right)$ then follows from the unrestricted estimation as contained in Theorem 2. As shown above the rates of convergence as well as the limits are unchanged for the coefficients corresponding to the non-stationary components of the regressors for the long VAR case compared to the finite VAR case.

Note that these results hold for general closed parameter space $\Theta$, thus including the unrestricted as well as the rank-reduced problem. This shows that we can always reduce the asymptotic analysis of the eigenvalue problems to a neighborhood of the true value as is done in P .

The first step in the proof of Theorem 4.1. of P consists in the investigation of the solutions to the equation $\left(\tilde{\beta}=\beta H+\beta_{1} H_{1}+\beta_{2} H_{2}\right.$, letting $B_{T}^{\prime}=\left[\begin{array}{c}\beta^{\prime} \\ T^{-1 / 2} \beta_{1}^{\prime} \\ T^{-3 / 2} \beta_{2}^{\prime}\end{array}\right]$ )

$$
B_{T}^{\prime} M_{22.1} B_{T}\left[\begin{array}{c}
H  \tag{A6}\\
T^{1 / 2} H_{1} \\
T^{3 / 2} H_{2}
\end{array}\right] \Lambda=B_{T}^{\prime} M_{20.1} M_{00.1}^{-1} M_{02.1} B_{T}\left[\begin{array}{c}
H \\
T^{1 / 2} H_{1} \\
T^{3 / 2} H_{2}
\end{array}\right] .
$$

Now Lemma A. 4 implies that the matrix $B_{T}^{\prime} M_{22.1} B_{T}$ on the left hand side converges to $\operatorname{diag}\left(\Sigma_{Y Y . U}, \int_{0}^{1} F_{+} F_{\dagger}^{\prime}\right) . B_{T}^{\prime} M_{20.1}=\left[\begin{array}{c}\Sigma_{Y 0 . U} \\ 0\end{array}\right]+O_{P}\left(h^{1 / 2} T^{-1 / 2}\right), M_{00.1}=\Sigma_{00 . U}+O_{P}\left(T^{-1 / 2}\right)$. Multiplying the equation by $h^{2}$ we obtain the limiting eigenvalue problem

$$
\left[\begin{array}{cc}
\tilde{\Sigma}_{Y Y . U} & O_{P}\left(T^{-1 / 2} h^{3 / 2}\right) \\
O_{P}\left(T^{-1 / 2} h^{3 / 2}\right) & h \int_{0}^{1} F_{+} F_{+}^{\prime}
\end{array}\right]\left[\begin{array}{c}
H \\
T^{1 / 2} H_{1} \\
T^{3 / 2} H_{2}
\end{array}\right] h \Lambda=\left[\begin{array}{cc}
\tilde{\Sigma}_{Y 0 . u} \Sigma_{00 . U}^{-1} \tilde{\Sigma}_{0 Y . U} & O_{P}\left(T^{-1 / 2} h^{5 / 2}\right) \\
O_{P}\left(T^{-1 / 2} h^{5 / 2}\right) & O_{P}\left(T^{-1} h^{3}\right)
\end{array}\right]\left[\begin{array}{c}
H \\
T^{1 / 2} H_{1} \\
T^{3 / 2} H_{2}
\end{array}\right] .
$$

equation
Therefore asymptotically the first $p-r$ eigenvalues of $h \Lambda$ are positive, the remaining ones tending to zero. Likewise the eigenvectors converge at the same speed as the matrices. Thus $H_{1}=$ $O_{P}\left(h^{5 / 2} / T\right), H_{2}=O_{P}\left(h^{5 / 2} / T^{2}\right)$ from which

$$
\beta^{\prime} M_{22.1} \beta H \Lambda H^{-1}=\beta^{\prime} M_{20.1} M_{00.1}^{-1} M_{02.1} \beta+O_{P}\left(h^{4} / T\right)
$$

and thus using (A.11)

$$
H \Lambda H^{-1}=\tilde{\Sigma}_{Y Y . U}^{-1} \tilde{\Sigma}_{Y 0 . U} \Sigma_{00 . U}^{-1} \tilde{\Sigma}_{0 Y . U} / h+O_{P}\left(h T^{-1 / 2}\right)=\alpha^{\prime} \Sigma_{00 . U}^{-1} \alpha \tilde{\Sigma}_{Y Y . U} / h+O_{P}\left(h T^{-1 / 2}\right)
$$

follows. Then as in P we have ${ }^{4}$

$$
M_{22.1} \underline{\tilde{\beta}}=M_{20.1}\left(\Sigma_{00 . U}^{-1} \tilde{\Sigma}_{0 Y . U}\left(h H \Lambda H^{-1}\right)^{-1}+O_{P}\left(h T^{-1 / 2}\right)\right)=M_{22.1} \beta+M_{2 \varepsilon .1} \Sigma_{\epsilon}^{-1} \alpha\left(\alpha^{\prime} \Sigma_{\epsilon}^{-1} \alpha\right)^{-1}+a_{1}
$$

where $a_{1}=M_{20.1} O_{P}\left(h^{2} T^{-1 / 2}\right)=o_{P}(1)$ and $\tilde{\beta}=\tilde{\beta} H^{-1}$. Then the remaining arguments on p .546 of P show that the asymptotic distribution of $\left(T \bar{\beta}_{1}, T^{2} \beta_{2}\right)^{\prime}(\underline{\tilde{\beta}}-\beta)$ is identical for the long VAR case as in the finite VAR case.

From these arguments the distribution of the likelihood ratio test of $H_{r}$ versus $H_{p}$ can be shown: Define $S_{1}(\lambda):=\lambda M_{22.1}-M_{20.1} M_{00.1}^{-1} M_{02.1}, A_{T}:=\left(\beta_{1}, T^{-1} \beta_{2}\right)$ and $\tilde{B}_{T}:=\left(\beta, A_{T}\right)=\left(\beta, \beta_{1}, T^{-1} \beta_{2}\right)$. Note that $\tilde{B}_{T}$ is of full rank, (11) is equivalent to $\left|\tilde{B}_{T}^{\prime} S_{1}(\lambda) \tilde{B}_{T}\right|=0$; that is,

$$
\left|\left(\begin{array}{c}
\beta^{\prime}  \tag{A7}\\
\beta_{1}^{\prime} \\
T^{-1} \beta_{2}^{\prime}
\end{array}\right) S_{1}(\lambda)\left(\beta, \beta_{1}, T^{-1} \beta_{2}\right)\right|=\left|\beta^{\prime} S_{1}(\lambda) \beta\right| \cdot\left|A_{T}^{\prime}\left(S_{1}(\lambda)-S_{1}(\lambda) \beta\left(\beta^{\prime} S_{1}(\lambda) \beta\right)^{-1} \beta^{\prime} S_{1}(\lambda)\right) A_{T}\right|=0
$$

Let $\delta_{1}=T \lambda$, so that for every $\delta_{1}$ we have that $\lambda \rightarrow 0$, as $T \rightarrow \infty$. By the above arguments we have that

$$
h^{2}\left|\beta^{\prime} S_{1}(\lambda) \beta\right|=\left|\delta_{1} \frac{h^{2}}{T} \beta^{\prime} M_{22.1} \beta-h^{2} \beta^{\prime} M_{20.1} M_{00.1}^{-1} M_{02.1} \beta\right| \xrightarrow{p}\left|-\tilde{\Sigma}_{Y 0 . U} \Sigma_{00 . U}^{-1} \tilde{\Sigma}_{0 Y . U}\right| \neq 0
$$

which has no zero root. Moreover, we have

$$
h A_{T}^{\prime} S_{1}(\lambda) \beta=h \lambda A_{T}^{\prime} M_{22.1} \beta-h A_{T}^{\prime} M_{20.1} M_{00.1}^{-1} M_{02.1} \beta=-A_{T}^{\prime} M_{20.1} \Sigma_{00 . U}^{-1} \tilde{\Sigma}_{0 \gamma . U}+o_{P}(1)
$$

which yields that

$$
\begin{aligned}
& \left|A_{T}^{\prime}\left(S_{1}(\lambda)-S_{1}(\lambda) \beta\left(\beta^{\prime} S_{1}(\lambda) \beta\right)^{-1} \beta^{\prime} S_{1}(\lambda)\right) A_{T}\right| \\
= & \left|\left(\delta_{1} \frac{1}{T} A_{T}^{\prime} M_{22.1} A_{T}-A_{T}^{\prime} M_{20.1} M_{00.1}^{-1} M_{02.1} A_{T}\right)-A_{T}^{\prime} S_{1}(\lambda) \beta\left(\beta^{\prime} S_{1}(\lambda) \beta\right)^{-1} \beta^{\prime} S_{1}(\lambda) A_{T}\right| \\
= & \left|\left(\delta_{1} \frac{1}{T} A_{T}^{\prime} M_{22.1} A_{T}\right)-A_{T}^{\prime} M_{20.1}\left(M_{00.1}^{-1}-\Sigma_{00 . U}^{-1} \tilde{\Sigma}_{0 Y . u}\left(\tilde{\Sigma}_{Y 0 . u} \Sigma_{00 . U}^{-1} \tilde{\Sigma}_{0 Y . U}\right)^{-1} \tilde{\Sigma}_{Y 0 . u} \Sigma_{00 . U}^{-1}+o_{P}(1)\right) M_{02.1} A_{T}\right| \\
= & \left|\left(\delta_{1} \frac{1}{T} A_{T}^{\prime} M_{22.1} A_{T}\right)-A_{T}^{\prime} M_{20.1}\left(\Sigma_{00 . U}^{-1}-\Sigma_{00 . U}^{-1} \alpha\left(\alpha \Sigma_{\epsilon}^{-1} \alpha\right)^{-1} \alpha^{\prime} \Sigma_{00 . U}^{-1}+o_{P}(1)\right) M_{02.1} A_{T}\right| \\
& \xrightarrow{d}\left|\delta_{1} \int_{0}^{1} F_{+} F_{\dagger}^{\prime}-\int_{0}^{1} F_{+} d W^{\prime} \alpha_{\perp}\left(\alpha_{\perp}^{\prime} \Sigma_{\epsilon} \alpha_{\perp}\right)^{-1} \alpha_{\perp}^{\prime} \int_{0}^{1} d W F_{\dagger}^{\prime}\right|=\left|\delta_{1} \int_{0}^{1} F_{+} F_{\dagger}^{\prime}-\left(\int_{0}^{1} F_{+} d W_{+}^{\prime}\right)\left(\int_{0}^{1} d W_{+} F_{+}^{\prime}\right)\right|
\end{aligned}
$$

where $W_{\dagger}=\left(\alpha_{\perp}^{\prime} \Sigma_{\varepsilon} \alpha_{\perp}\right)^{-1 / 2} \alpha_{\perp}^{\prime} W$. Thus, the smallest $(p-r)$ solutions of (11) converge in distribution to the solutions of $\left|\delta_{1} \int_{0}^{1} F_{+} F_{\dagger}^{\prime}-\left(\int_{0}^{1} F_{\dagger} d W_{\dagger}^{\prime}\right)\left(\int_{0}^{1} d W_{+} F_{\dagger}^{\prime}\right)\right|=0$, which implies that the test statistic $Q_{r}$ has the following limiting distribution,

$$
Q_{r}=\sum_{i=r+1}^{p} \delta_{1, i}+o_{P}(1) \xrightarrow{d} \operatorname{tr}\left(\left(\int_{0}^{1} d W_{+} F_{\dagger}^{\prime}\right)\left(\int_{0}^{1} F_{\dagger} F_{\dagger}^{\prime}\right)^{-1}\left(\int_{0}^{1} F_{\dagger} d W_{\dagger}^{\prime}\right)\right) .
$$

For the second stage the arguments are very similar. The eigenvalue problem solved here is the following:

$$
\tilde{\tilde{\beta}}_{\perp}^{\prime} M_{11 . \tilde{\tilde{\beta}}} \tilde{\tilde{\beta}}_{\perp} \tilde{\eta} Y=\tilde{\tilde{\beta}}_{\perp}^{\prime} M_{1 \tilde{\alpha}_{\perp} \cdot \tilde{\beta}} M_{\tilde{\alpha}_{\perp} \tilde{\alpha}_{\perp} \cdot \tilde{\beta}}^{-1} M_{\tilde{\alpha}_{\perp} 1 . \tilde{\beta}} \overline{\tilde{\beta}}_{\perp} \tilde{\eta} .
$$

[^28]This formula uses $\tilde{\alpha}_{\perp}$, the ortho-complement of

$$
\tilde{\alpha}=M_{02.1} \tilde{\beta}\left(\tilde{\beta}^{\prime} M_{22.1} \tilde{\beta}\right)^{-1}
$$

From the above results noting that $h \tilde{\beta}^{\prime} M_{22.1} \tilde{\beta} \rightarrow \tilde{\Sigma}_{Y Y . U}$ and $h M_{02.1} \tilde{\beta} \rightarrow \alpha \tilde{\Sigma}_{Y Y . U}$ according to Lemma A. 4 we have $\tilde{\alpha} \rightarrow \alpha$. Considering the order of convergence we obtain $\tilde{\alpha}-\alpha=O_{P}\left(h T^{-1 / 2}\right)$. As in P this implies $\tilde{\alpha}_{\perp}-\alpha_{\perp}=O_{P}\left(h T^{-1 / 2}\right)$. Using $\tilde{\beta}-\beta=O_{P}\left(h^{5 / 2} / T\right)$ from stage 1 one observes that in the eigenvalue problem estimates can be replaced by true quantities introducing an error of order $o_{P}\left(h T^{-1 / 2}\right)$ :

$$
\bar{\beta}_{\perp}^{\prime} M_{11 . \beta} \tilde{\beta}_{1} Y=\bar{\beta}_{\perp}^{\prime} M_{1 \alpha_{\perp} \cdot \beta} M_{\alpha_{\perp} \alpha_{\perp} \cdot \beta}^{-1} M_{\alpha_{\perp} 1 . \beta} \tilde{\beta}_{1}+o_{P}\left(h T^{-1 / 2}\right)
$$

Then as in P consider $\tilde{\beta}_{1}=\beta H+\beta_{1} H_{1}+\bar{\beta}_{2} H_{2}$, reusing the symbols $H, H_{1}, H_{2}$ here for $\tilde{\beta}_{1}$ in place of $\tilde{\beta}$ as before. Identical arguments as around (A6) show that $H_{1}=O_{P}(1)$ and $H_{2}=O_{P}\left(h^{2} / T\right)$. Then combining the arguments around (A6) with the developments in P, p. 546 and 547 we obtain (A.21) of P:

$$
\bar{\beta}_{\perp}^{\prime} M_{11 . \beta}\left(\underline{\tilde{\beta}}_{1}-\beta_{1}\right)=\bar{\beta}_{\perp}^{\prime} M_{1 \varepsilon . \beta} \alpha_{\perp} \Sigma_{\alpha_{\perp} \alpha_{\perp}}^{-1} \zeta\left(\zeta^{\prime} \Sigma_{\alpha_{\perp} \alpha_{\perp}}^{-1} \zeta\right)^{-1}+o_{P}(1) .
$$

The rest of the proof of (4.3a) and (4.3b) of P follows as in P.
With respect to the second likelihood ratio test consider

$$
\tilde{S}_{2}(\rho)=\rho \overline{\tilde{\beta}}_{\perp}^{\prime} M_{11 . \tilde{\beta}} \overline{\tilde{\beta}}_{\perp}-\overline{\tilde{\beta}}_{\perp}^{\prime} M_{1 \tilde{\alpha}_{\perp} \cdot \tilde{\beta}} M_{\tilde{\alpha}_{\perp} \tilde{\tilde{L}}_{\perp} \cdot \tilde{\beta}}^{-1} M_{\tilde{\alpha}_{\perp} 1 . \tilde{\beta}} \overline{\tilde{\beta}}_{\perp} .
$$

The results above imply that $\tilde{S}_{2}(\rho)$ has uniformly in $|\rho|<C$ (for every $0<C<\infty$ ) distance to $S_{2}(\rho)$ of order $O_{P}\left(h T^{-1 / 2}\right)$ where

$$
S_{2}(\rho)=\rho \bar{\beta}_{\perp}^{\prime} M_{11 . \beta} \bar{\beta}_{\perp}-\bar{\beta}_{\perp}^{\prime} M_{1 \alpha_{\perp} \cdot \beta} M_{\alpha_{\perp} \alpha_{\perp} \cdot \beta}^{-1} M_{\alpha_{\perp} 1 . \beta} \bar{\beta}_{\perp}
$$

Note that since $\left(\eta, \eta_{\perp}\right)$ is of full rank, (16) is equivalent to

$$
\begin{equation*}
\left|\binom{\eta^{\prime}}{\eta_{\perp}^{\prime}} S_{2}(\rho)\left(\eta, \eta_{\perp}\right)\right|=\left|\eta^{\prime} S_{2}(\rho) \eta\right| \cdot\left|\eta_{\perp}^{\prime}\left(S_{2}(\rho)-S_{2}(\rho) \eta\left(\eta^{\prime} S_{2}(\rho) \eta\right)^{-1} \eta^{\prime} S_{2}(\rho)\right) \eta_{\perp}\right|=0 . \tag{A8}
\end{equation*}
$$

Let $\delta_{2}=T \rho$, so that $\rho \rightarrow 0$, as $T \rightarrow \infty$. As above it can be seen that

$$
h^{2}\left|\eta^{\prime} S_{2}\left(\frac{\delta_{2}}{T}\right) \eta\right|=h^{2}\left|\frac{\delta_{2}}{T} \beta_{1}^{\prime} M_{11 . \beta} \beta_{1}-\beta_{1}^{\prime} M_{1 \bar{\alpha}_{\perp} \cdot \beta} M_{\bar{\alpha}_{\perp} \bar{\alpha}_{\perp} \cdot \beta}^{-1} M_{\bar{\alpha}_{\perp} 1 \cdot \beta} \beta_{1}\right| \xrightarrow{p}\left|-\tilde{\Sigma}_{U 0} \alpha_{\perp}\left(\alpha_{\perp}^{\prime} \Sigma_{00} \alpha_{\perp}\right)^{-1} \alpha_{\perp}^{\prime} \tilde{\Sigma}_{0 U}\right| \neq 0 .
$$

This shows that the $s$ larger roots of $S_{2}(\rho)$ tend to zero slower than $O(1 / T)$. Moreover, we have

$$
h \eta_{\perp}^{\prime} S_{2}\left(\frac{\delta_{2}}{T}\right) \eta=h\left(\frac{\delta_{2}}{T} \beta_{2}^{\prime} M_{11 . \beta} \beta_{1}-\beta_{2}^{\prime} M_{1 \bar{\alpha}_{\perp} \cdot \beta} M_{\bar{\alpha}_{\perp} \bar{\alpha}_{\perp} \cdot \beta}^{-1} M_{\bar{\alpha}_{\perp} 1 . \beta} \beta_{1}\right)=-\beta_{2}^{\prime} M_{1 \bar{\alpha}_{\perp} \cdot \beta}\left(\alpha_{\perp}^{\prime} \Sigma_{00} \alpha_{\perp}\right)^{-1} \alpha_{\perp}^{\prime} \tilde{\Sigma}_{0 U}+o_{P}(1),
$$

which yields that (using $\left.P_{M}:=\left(\alpha_{\perp}^{\prime} \Sigma_{00} \alpha_{\perp}\right)^{-1} \alpha_{\perp}^{\prime} \tilde{\Sigma}_{0 U}\left(\tilde{\Sigma}_{U 0} \alpha_{\perp}\left(\alpha_{\perp}^{\prime} \Sigma_{00} \alpha_{\perp}\right)^{-1} \alpha_{\perp}^{\prime} \tilde{\Sigma}_{0 U}\right)^{-1} \tilde{\Sigma}_{U 0} \alpha_{\perp}\left(\alpha_{\perp}^{\prime} \Sigma_{00} \alpha_{\perp}\right)^{-1}\right)$

$$
\begin{aligned}
&\left.\left\lvert\, \eta_{\perp}^{\prime}\left(S_{2} \frac{\delta_{2}}{T}\right)-S_{2}\left(\frac{\delta_{2}}{T}\right) \eta\left(\eta^{\prime} S_{2}\left(\frac{\delta_{2}}{T}\right) \eta\right)^{-1} \eta^{\prime} S_{2}\left(\frac{\delta_{2}}{T}\right)\right.\right) \eta_{\perp} \mid \\
&=\left|\left(\delta_{2} \frac{1}{T} \beta_{2}^{\prime} M_{11 . \beta} \beta_{2}-\beta_{2}^{\prime} M_{1 \bar{\alpha}_{\perp} . \beta} M_{\bar{\alpha}_{\perp} \bar{\alpha}_{\perp} \cdot \beta}^{-1} M_{\bar{\alpha}_{\perp} 1 . \beta} \beta_{2}\right)-h \eta_{\perp}^{\prime} S_{2}\left(\frac{\delta_{2}}{T}\right) \eta\left(h^{2} \eta^{\prime} S_{2}\left(\frac{\delta_{2}}{T}\right) \eta\right)^{-1} h \eta^{\prime} S_{1}\left(\frac{\delta_{2}}{T}\right) \eta_{\perp}\right| \\
&=\left|\left(\delta_{2} \frac{1}{T} \beta_{2}^{\prime} M_{11 . \beta} \beta_{2}-\beta_{2}^{\prime} M_{1 \alpha_{\perp} \cdot \beta}\left(\alpha_{\perp}^{\prime} \Sigma_{00} \alpha_{\perp}\right)^{-1} M_{\alpha_{\perp} 1 . \beta} \beta_{2}\right)+\beta_{2}^{\prime} M_{1 \alpha_{\perp} . \beta} P_{M} M_{\alpha_{\perp} 1 . \beta} \beta_{2}\right|+o_{P}(1) \\
& \xrightarrow{d}\left|\delta_{2} \int_{0}^{1} B_{3} B_{3}^{\prime}-\int_{0}^{1} B_{3} d W^{\prime} \alpha_{2}\left(\alpha_{2}^{\prime} \Sigma_{\epsilon} \alpha_{2}\right)^{-1} \alpha_{2}^{\prime} \int_{0}^{1} d W B_{3}^{\prime}\right|=\left|\delta_{2} \int_{0}^{1} B_{3} B_{3}^{\prime}-\left(\int_{0}^{1} B_{3} d W_{2}^{\prime}\right)\left(\int_{0}^{1} d W_{2} B_{3}^{\prime}\right)\right|
\end{aligned}
$$

using the results of Lemma A. 5 of P. and (A.18) of Paruolo (1996) as an expression for

$$
\left(\alpha_{\perp}^{\prime} \Sigma_{00} \alpha_{\perp}\right)^{-1}-P_{M}
$$

where $W_{2}=\left(\alpha_{2}^{\prime} \Sigma_{\epsilon} \alpha_{2}\right)^{-1 / 2} \alpha_{2}^{\prime} W$.
Thus, the smallest $(p-r-s)$ solutions of (16) converge in distribution to the solutions of

$$
\left|\delta_{2} \int_{0}^{1} B_{3} B_{3}^{\prime}-\left(\int_{0}^{1} B_{3} d W_{2}^{\prime}\right)\left(\int_{0}^{1} d W_{2} B_{3}^{\prime}\right)\right|=0
$$

which shows that the test statistic $Q_{r, s}$ has the following limiting distribution,

$$
Q_{r, s}=\sum_{i=s+1}^{p-r} \delta_{2, i}+o_{P}(1) \xrightarrow{d} \operatorname{tr}\left(\int_{0}^{1} d W_{2} B_{3}^{\prime}\left(\int_{0}^{1} B_{3} B_{3}^{\prime}\right)^{-1} \int_{0}^{1} B_{3} d W_{2}^{\prime}\right) .
$$

It follows also that the sum $S_{r, s}=Q_{s}+Q_{r, s}$ converges in distribution showing (C).
The rest of the proof of relations (4.3a, b) of $P$ follow exactly as in P. In P (4.4) the order of convergence is replaced by $o_{P}\left(T^{-1}\right)$, in (4.5) the error term can be shown to be $o_{P}\left(T^{-1 / 2}\right)$ and in (4.6) instead of the term $O_{P}\left(T^{-2}\right)$ we achieve $o_{P}(1)$.

These terms show consistency for $\tilde{\beta}, \tilde{\eta}$. Using the results of Lemma A. 4 of $P$ then consistency for $\tilde{\alpha}, \tilde{\zeta}$ follow.

Following the proof of Theorem 4.2. on pp. 548+549 of P we can show consistency for $\tilde{\psi}$ of P . The only changes refer to the orders of convergence where our setting introduces orders of $h$ into the arguments. Jointly this proves consistency of $\tilde{\Psi}$ and $\tilde{\Gamma}$. Consistency for the coefficients to the stationary terms $\Delta^{2} y_{t-j}$ follows as usual from the consistency of the estimates for the coefficients to non-stationary regressors. This completes the proof of (D).

With respect to (E) note that the results above show that the asymptotics for the two eigenvalue problems to be solved converge to the same quantities as in the finite VAR case. This shows that the results of P in this respect hold also in the case of long VARs.

Finally for the matrices $\Pi_{j}$ note that Theorem 4.3. of $P$ shows that the asymptotic distribution for all quantities corresponding to stationary regressors are identical for every super-consistent estimator for the coefficients to the non-stationary components.

## Appendix E. Proof of Theorem 4

From Theorem 3 it follows that $\hat{\Phi}=\hat{\alpha} \hat{\beta}^{\prime} \rightarrow \Phi, \hat{\Psi} \rightarrow \Psi, \hat{\Pi}_{j} \rightarrow \Pi_{j}, j=1,2, \ldots, 2 f-1$. Therefore the Hankel matrix of impulse response coefficients $\hat{\Pi}_{j}$ converges to the Hankel matrix corresponding to the $\Pi_{j}^{\prime} \mathrm{s}$. As $(\bar{A}, B)$ is controllable, $(A, B, C)$ is minimal and $\bar{A}$ is nonsingular according to the assumptions, this Hankel matrix has rank $n$. This implies that the stochastic realisation algorithm of Appendix F provides consistent estimates $(\hat{A}, \hat{B}, \hat{D}) \rightarrow(\bar{A}, B, D)$. This implies

$$
\hat{a}(z)=(1-z)^{2} I_{p}-\hat{\Phi} z-\hat{\Psi} z(1-z)-(1-z)^{2} z \hat{D}\left(I_{n}-z \hat{A}\right)^{-1} \hat{B} \rightarrow a(z)
$$

For details see Appendix F.
$\hat{a}(z)$ does not necessarily correspond to a rational transfer function of order $n$. It does so, however, if the additional restrictions (22) hold. Step 3 and 4 of the proposed algorithm achieve this. Here step 3 ascertains that solutions to the third equation exist. The second equation explicitly provides a solution $\bar{\alpha}_{\perp}$ for given $C_{+}$. This solution not necessarily is of full row rank. As in the limit this is the case, it also holds for large enough $T$. The first equation always admits solutions. Thus for large enough $T$ the set of all solutions is defined by polynomial restrictions. Adding the least squares distance to the estimated impulse response sequence then leads to a quadratic problem under non-linear differentiable constraints, which in the limit has a unique solution. Thus the solution is unique for large enough $T$.

Consistency of the estimates in combination with continuity of the solution of step 4 implies consistency for the system $(\hat{A}, \hat{B}, \hat{C})$. This implies consistency for the inverse system $(\hat{A}, \hat{B}, \hat{C})$ in the sense of converging impulse response coefficients and hence consistency for the transfer function estimator in the pointwise topology. The fulfillment of restrictions (22) ensures the structure of the corresponding matrix $\hat{A}$ according to state space unit root structure $((0,(c, c+d)))$.

## Appendix F. Stochastic Realization Using Overlapping Echelon Forms

This section describes the approximate realization of the first $f$ coefficients $G_{j}, j=1, \ldots, 2 f$ of an impulse response sequence using a rational transfer function of order $n$ where $f \geq n$. More details can be found in Section 2.6. of Hannan and Deistler (1988).

Define the Hankel matrix

$$
\mathcal{H}_{f, f}=\left[\begin{array}{ccccc}
G_{1} & G_{2} & G_{3} & \ldots & G_{f} \\
G_{2} & G_{3} & \ldots & & \\
G_{3} & \ldots & & & \vdots \\
\vdots & & & & \vdots \\
G_{f} & G_{f+1} & \ldots & & G_{2 f-1}
\end{array}\right]=\left[\begin{array}{c}
h(1,1) \\
h(1,2) \\
\vdots \\
h(1, p) \\
h(2,1) \\
\vdots \\
h(f, p)
\end{array}\right] .
$$

Here $h(i, j)$ denotes the $j$-th row in the $i$-th block row. Let $\alpha=\left(n_{1}, \ldots, n_{p}\right)$ define a nice selection of rows ${ }^{5}$ of $\mathcal{H}$ such that $\mathcal{H}_{\alpha} \in \mathbb{R}^{n \times f p}$, the submatrix of $\mathcal{H}$ containing the rows $h(i, j), i \leq n_{j}$, is of full row rank. If the impulse response corresponds to a transfer function of order at least $n$ there exists such a nice selection $\alpha$. Finally let $\mathcal{H}_{\alpha+1} \in \mathbb{R}^{n \times f p}$ denote the matrix $\mathcal{H}_{\alpha}$ shifted down one block row (that is in each row where $\mathcal{H}_{\alpha}$ contains $h(i, j), \mathcal{H}_{\alpha+1}$ contains $\left.h(i+1, j)\right)$.

Then it is derived in Hannan and Deistler (1988), Theorem 2.6.2. that if $G_{j}$ corresponds to a transfer function $k(z)=\sum_{j=1}^{\infty} G_{j} z^{-j}$ of order exactly $n$ such that the corresponding $\mathcal{H}_{\alpha}$ is formed using a nice selection, then a system $(A, B, C)$ can be defined using the following formulas

$$
A \mathcal{H}_{\alpha}=\mathcal{H}_{\alpha+1}, \quad B=\mathcal{H}_{\alpha}\left[\begin{array}{c}
I_{p}  \tag{A9}\\
0
\end{array}\right], \quad C \mathcal{H}_{\alpha}=\left[\begin{array}{llll}
G_{1} & G_{2} & \ldots & G_{f}
\end{array}\right]
$$

such that $G_{j}=C A^{j-1} B, j=1,2, \ldots$..
If the order of the transfer function is larger than $n$, then the equations for $A$ and $C$ can be solved using least squares. If a sequence of impulse responses $\hat{G}_{j} \rightarrow G_{j}, j=1, \ldots, 2 f-1$, and the limit $G_{j}$ corresponds to a transfer function where the rank of $\mathcal{H}_{\alpha}$ equals $n$, it is obvious that the resulting systems $(\hat{A}, \hat{B}, \hat{C}) \rightarrow(A, B, C)$ since in this case the least squares solution depends continuously on the matrix $\mathcal{H}$.

## References

Banerjee, Anindya, Lynne Cockerell, and Bill Russell. 2001. An I(2) analysis of inflation and the markup. Journal of Applied Econometrics 16: 221-40. [CrossRef]
Bauer, Dietmar, and Alex Maynard. 2012. Persistence-robust surplus-lag Granger causality testing. Journal of Econometrics 169: 293-300. [CrossRef]
Bauer, Dietmar, and Martin Wagner. 2004. Autoregressive Approximations to MFI(1) Processes. Working Paper No. 174, Reihe Ökonomie/Economics Series, Institut für Höhere Studien (IHS), Vienna, Austria.

[^29]Bauer, Dietmar, and Martin Wagner. 2012. A state space canonical form for unit root processes. Econometric Theory 28: 1313-49. [CrossRef]
Berk, Kenneth N. 1974. Consistent autoregressive spectral estimates. The Annals of Statistics 2: 489-502. [CrossRef]
Boswijk, H. Peter, and Jurgen A. Doornik. 2004. Identifying, estimating and testing restricted cointegrated systems: An overview. Statistica Neerlandica 58: 440-65. [CrossRef]
Boswijk, H. Peter, and Paolo Paruolo. 2017. Likelihood ratio tests of restrictions on common trends loading matrices in I(2) VAR systems. Econometrics 5: 28. [CrossRef]
Chan, Ngai Hang, and Ching Zong Wei. 1988. Limiting distributions of least squares estimates of unstable autoregressive processes. The Annals of Statistics 16: 367-401. [CrossRef]
Dolado, Juan J., and Helmut Lütkepohl. 1996. Making Wald tests work for cointegrated VAR systems. Econometric Reviews 15: 369-86. [CrossRef]
Engle, Robert F., and Clive W.J. Granger. 1987. Co-integration and error correction: Representation, estimation, and testing. Econometrica 55: 251-76. [CrossRef]
Georgoutsos, Dimitris A., and Georgios P. Kouretas. 2004. A Multivariate I (2) cointegration analysis of German hyperinflation. Applied Financial Economics 14: 29-41. [CrossRef]
Hannan, Edward James, and Manfred Deistler. 1988. The Statistical Theory of Linear Systems. New York: John Wiley.
Hannan, Edward James, and Laimonis Kavalieris. 1986. Regression, autoregression models. Journal of Time Series Analysis 7: 27-49. [CrossRef]
Inoue, Atsushi, and Lutz Kilian. 2020. The uniform validity of impulse response inference in autoregressions. Journal of Econometrics 215: 450-72. [CrossRef]
Johansen, Søren, and Helmut Lütkepohl. 2005. A note on testing restrictions for the cointegration parameters of a VAR with I(2) variables. Econometric Theory 21: 653-58. [CrossRef]
Johansen, Søren, Katarina Juselius, Roman Frydman, and Michael Goldberg. 2007. Testing hypotheses in an I (2) model with applications to the persistent long swings in the Dmk/\$ rate. Journal of Econometrics 158: 1-35. [CrossRef]
Johansen, Søren. 1992a. A representation of vector autoregresive processes integrated of order 2. Econometric Theory 8: 188-202. [CrossRef]
Johansen, Søren. 1992b. Testing weak exogeneity and the order of cointegration in UK money demand data. Journal of Policy Modeling 14: 313-34. [CrossRef]
Johansen, Søren. 1995. Likelihood-Based Inference in Cointegrated Vector Auto-Regressive Models. Oxford: Oxford University Press.
Johansen, Søren. 1997. Likelihood analysis of the I(2) model. Scandinavian Journal of Statistics 24: 433-62. [CrossRef]
Juselius, Katarina, and Katrin Assenmacher. 2017. Real exchange rate persistence and the excess return puzzle: The case of Switzerland versus the US. Journal of Applied Econometrics 32: 1145-55. [CrossRef]
Juselius, Katarina, and Josh R. Stillwagon. 2018. Are outcomes driving expectations or the other way around? An I(2) CVAR analysis of interest rate expectations in the dollar/pound market. Journal of International Money and Finance 83: 93-105. [CrossRef]
Juselius, Katarina. 1994. On the duality between long-run relations and common trends in the $i(1)$ versus (2) model. an application to aggregate money holdings. Econometric Reviews 13: 151-78. [CrossRef]
Juselius, Katarina. 2006. The Cointegrated VAR Model. Oxford: Oxford University Press.
Kurita, Takamitsu, Heino Bohn Nielsen, and Anders Rahbek. 2011. An I(2) cointegration model with piecewise linear trends. Econometrics Journal 14: 131-55. [CrossRef]
Kurita, Takamitsu. 2012. Likelihood-based inference for weak exogeneity in $I(2)$ cointegrated VAR models. Econometric Reviews 31: 325-60. [CrossRef]
Lewis, Richard, and Gregory C. Reinsel. 1985. Prediction of multivariate time series by autoregressive model fitting. Journal of Multivariate Analysis 16: 393-411. [CrossRef]
Lütkepohl, Helmut, and Holger Claessen. 1997. Analysis of cointegrated VARMA processes. Journal of Econometrics 80: 223-29. [CrossRef]
Lütkepohl, Helmut, and Pentti Saikkonen. 1997. Impulse response analysis in infinite order cointegrated vector autoregressive processes. Journal of Econometrics 81: 127-57. [CrossRef]
Mosconi, Rocco, and Paolo Paruolo. 2013. Identification of Cointegrating Relations in I(2) Vector Autoregressive Models. Rome: PRIN Workshop Forecasting Economic and Financial Time Series, Research Publications at Politecnico di Milano, pp. 1-35.

Mosconi, Rocco, and Paolo Paruolo. 2017. Identification conditions in simultaneous systems of cointegrating equations with integrated variables of higher order. Journal of Econometrics 198: 271-76. [CrossRef]
Nielsen, Heino Bohn, and Anders Rahbek. 2007. The likelihood ratio test for cointegration ranks in the I(2) model. Econometric Theory 23: 615-37. [CrossRef]
Palma, Wilfredo, and Pascal Bondon. 2003. On the eigenstructure of generalized fractional processes. Statistics and Probability Letters 65: 93-101. [CrossRef]
Paruolo, Paolo, and Anders Rahbek. 1999. Weak exogeneity in I(2) VAR systems. Journal of Econometrics 93: 281-308. [CrossRef]
Paruolo, Paolo. 1994. The role of the drift in I(2) systems. Journal of the Italian Statistical Society 3: 93-123. [CrossRef]
Paruolo, Paolo. 1996. On the determination of integration indices in $\mathrm{I}(2)$ systems. Journal of Econometrics 72: 313-56. [CrossRef]
Paruolo, Paolo. 2000. Asymptotic efficiency of the two stage estimator in I(2) systems. Econometric Theory 16: 524-50. [CrossRef]
Paruolo, Paolo. 2006. Common trends and cycles in I(2) VAR systems. Journal of Econometrics 132: 143-68. [CrossRef]
Rahbek, Anders, Hans Christian Kongsted, and Clara Jørgensen. 1999. Trend stationarity in the I(2) cointegration model. Journal of Econometrics 90: 265-289. [CrossRef]
Saikkonen, Pentti, and Helmut Lütkepohl. 1996. Infinite-order cointegrated vector autoregressive processes : estimation and inference. Econometric Theory 12: 814-44. [CrossRef]
Saikkonen, Pentti, and Ritva Luukkonen. 1997. Testing cointegration in infinite order vector autoregressive processes. Journal of Econometrics 81: 93-126. [CrossRef]
Saikkonen, Pentti. 1991. Asymptotically efficient estimation of cointegration regressions. Econometric Theory 7: 1-21. [CrossRef]
Saikkonen, Pentti. 1992. Estimation and testing of cointegrated systems by an autoregressive approximation. Econometric Theory 8: 1-27. [CrossRef]
Sims, Christopher A., James H. Stock, and Mark W. Watson. 1990. Inference in linear time series models with some unit roots. Econometrica 58: 113-44. [CrossRef]
Stillwagon, Josh R. 2018. Are risk premia related to real exchange rate swings? Evidence from I(2) CVARs with survey expectations. Macroeconomic Dynamics 22: 255-78. [CrossRef]

# Article <br> The Stochastic Stationary Root Model 

Andreas Hetland<br>Department of Economics, University of Copenhagen, 1353 Copenhagen K, Denmark; lxs601@ku.dk; Tel.: +45-20422606

Received: 31 March 2018; Accepted: 13 August 2018; Published: 21 August 2018


#### Abstract

We propose and study the stochastic stationary root model. The model resembles the cointegrated VAR model but is novel in that: (i) the stationary relations follow a random coefficient autoregressive process, i.e., exhibhits heavy-tailed dynamics, and (ii) the system is observed with measurement error. Unlike the cointegrated VAR model, estimation and inference for the SSR model is complicated by a lack of closed-form expressions for the likelihood function and its derivatives. To overcome this, we introduce particle filter-based approximations of the log-likelihood function, sample score, and observed Information matrix. These enable us to approximate the ML estimator via stochastic approximation and to conduct inference via the approximated observed Information matrix. We conjecture the asymptotic properties of the ML estimator and conduct a simulation study to investigate the validity of the conjecture. Model diagnostics to assess model fit are considered. Finally, we present an empirical application to the 10-year government bond rates in Germany and Greece during the period from January 1999 to February 2018.


Keywords: cointegration; particle filtering; random coefficient autoregressive model; state space model; stochastic approximation

JEL Classification: C15; C32; C51; C58

## 1. Introduction

In this paper, we introduce the multivariate stochastic stationary root (SSR) model. The SSR model is a nonlinear state space model, which resembles the Granger-Johansen representation of the cointegrated vector autoregressive (CVAR) model, see inter alia Johansen (1996) and Juselius (2007). The SSR model decomposes a $p$-dimensional observation vector into $r$ stationary components and $p-r$ nonstationary components, which is similar to the CVAR model. However, the roots of the stationary components are allowed to be stochastic; hence the name 'stochastic stationary root'. The stationary and nonstationary dynamics of the model are observed with measurement error, which in this model prohibits close-form expressions for e.g., the log-likelihood, sample score and observed Information matrix. Likelihood-based estimation and inference therefore calls for non-standard methods.

Although the SSR model resembles the CVAR model, it is differentiated by its ability to characterize heavy-tailed dynamics in the stationary component. Heavy-tailed dynamics, and other types of nonlinear dependencies, are not amenable to analysis with the CVAR model, which has prompted work into nonlinear alternatives, see inter alia Bohn Nielsen and Rahbek (2014), Kristensen and Rahbek (2013), Kristensen and Rahbek (2010), and Bec et al. (2008). Similarly, cointegration in the state space setting has been considered in term of the common stochastic trend (CST) model by Chang et al. (2009) as well as the CVAR model with measurement errors by Bohn Nielsen (2016). Additionally, the SSR model is also related to the stochastic unit root literature, see inter alia Granger and Swanson (1997), Leybourne and McCabe (1996), Lieberman and Phillips (2014), Lieberman and Phillips (2017), McCabe and Tremayne (1995), and McCabe and Smith (1998). Relevant empirical applications where the SSR model could potentially provide a better fit than the CVAR model include, but are not limited to, (i) log-prices of assets that exhibit random walk behavior in the levels and heavy-tailed error-correcting dynamics in the
no-arbitrage relations, and (ii) interest rates for which the riskless rate exhibits random walk-type dynamics and the risk premia undergo periods of high levels and high volatility.

The stationary and nonstationary components of the SSR model are treated as unobserved processes, and consequently need to be integrated out in order to compute the log-likelihood function and its derivatives. Due to the nonlinearity of the model, this cannot be accomplished analytically. We appeal to the incomplete data framework and the simulation-based approach known as particle filtering to approximate the log-likelihood function, sample score and observed Information matrix. See inter alia Gordon et al. (1993), Doucet et al. (2001), Cappé et al. (2005), and Creal (2012) for an overview of the particle filtering literature. Moreover, we rely on stochastic approximation methods to obtain the maximum likelihood (ML) estimator, see Poyiadjis et al. (2011). Summarizing, the main contributions of this paper are to
i introduce and study the SSR model, and
ii propose a method for approximate frequentist estimation and inference.
It is beyond the scope of this paper to provide a complete proof of the asymptotic properties of the ML estimator. The study of the asymptotic properties of the ML estimator in general state space models, such as the SSR model, is an emerging area of research. Most existing results rely on compactness of the state space, which excludes the SSR model and is generally restrictive. For results in this direction, see e.g., Olsson and Rydén (2008) who derive consistency and asymptotic normality for the ML estimator by discretizing the parameter space. Douc et al. (2011) have shown consistency of the ML estimator without assuming compactness, but the regularity conditions are nonetheless too restrictive to encompass the SSR model. Instead of providing a complete proof of the asymptotic properties of the ML estimator, we conjecture the asymptotic properties of the derivatives of the log-likelihood function. We base the conjecture on known properties of models that are closely related to the SSR model, and corroborate it by a simulation study. Given the conjecture holds, it allows us to establish the asymptotic properties of the ML estimator. We leave proving the conjecture for future work, and focus in this paper on developing methods for approximate frequentist estimation and inference.

The rest of the paper is organized as follows. We introduce the SSR model in Section 2, and study some properties of the process in Section 3. In Section 4 we introduce likelihood-based estimation and inference for the unknown model parameter. In Section 5 we introduce the incomplete data framework. In Section 6 we introduce the particle filter-based approximations to the log-likelihood function, sample score and Information matrix. In Section 7 we propose how to approximate the ML estimator and classic standard errors. In Section 8 we consider model diagnostics. In Section 9 we conduct a simulation study of the asymptotic distribution of the ML estimator. In Section 10 we apply the SSR model to monthly observations of 10-year government bond rates in Germany and Greece from January 1999 to February 2018. We conclude in Section 11. All proofs have been relegated to Appendix B, while Appendix A contains various auxiliary results.

Notation-wise, we adopt the convention that the 'blackboard bold' typeface, e.g., $\mathbb{E}$, denotes operators, and the 'calligraphy' typeface, e.g., $\mathcal{X}$, denotes sets. We thus let $\mathcal{R}$ and $\mathcal{N}$ denote the real and natural numbers, respectively. For any matrix $A$, we denote by $|A|$ the determinant, by $\|A\|=\sqrt{\operatorname{tr}\left(A^{\prime} A\right)}$ the Euclidean norm, and by $\rho(A)$ the spectral radius. For some positive definite matrix $A$, we let $A^{1 / 2}$ denote the lower triangular Cholesky decomposition. For some function $f: \mathcal{R}^{d_{z}} \mapsto \mathcal{R}^{d_{f}}$, let $\partial f(z) / \partial z$ denote the derivative of $f(z)$ with respect to $z$. For some stochastic variable $z \in \mathcal{R}^{d_{z}}$ with Gaussian distribution with mean $\mu$ and covariance $\Sigma$, let $N(z ; \mu, \Sigma)$ denote the Gaussian probability density function evaluated at $z$. We let $p(z)$ denote the probability density of stochastic variable $z \in \mathcal{R}^{d_{z}}$ with respect to the $d_{z}$-dimensional Lebesgue measure $m$, while $p(\mathrm{~d} z)=p(z) \mathrm{d} m$ denotes the corresponding probability measure. Additionally, the letter ' p ' is generic notation for probability density functions and measures induced by the model defined in (1)-(3) below. The 'bold' typeface, e.g., $p$, is generic notation for analytically intractable quantities,
in the sense of having no closed-form expression. Finally, we denote a sequence of $n \in \mathcal{N}_{+}$real $d_{z}$-dimensional vectors by $z_{1: n}:=\left[\begin{array}{lll}z_{1}^{\prime} & \ldots & z_{n}^{\prime}\end{array}\right]^{\prime} \in \mathcal{R}^{n \times d_{z}}$.

## 2. The Model

The structure of the SSR model is similar to the Granger-Johansen representation of the CVAR model, cf. Johansen (1996, chp. 4), but departs from it in two respects. First, the stationary component is a random coefficient autoregressive process, cf. e.g., Feigin and Tweedie (1985), rather than an autoregressive process. Second, the stationary and nonstationary components are observed with measurement error. This makes the SSR model is a state space model, whereas the CVAR model is observation-driven. In addition to resembling the CVAR model, the SSR model constitutes an extension of the CST model, cf. Chang et al. (2009). However, while the CST model is a linear Gaussian state space model, the SSR model is a nonlinear Gaussian state space model as it allows the stationary component to be a random coefficient autoregressive process.

Formally, we consider the observable $p$-dimensional discrete time vector process $y_{t}$, for $t=1,2, \ldots, T$ given by,

$$
\begin{gather*}
y_{t}=C\left(y_{0}\right)+B \sum_{i=1}^{t} \eta_{i}+A \xi_{t}+u_{t}  \tag{1}\\
\xi_{t}=\mu+\Phi_{t} \xi_{t-1}+v_{t} \tag{2}
\end{gather*}
$$

for fixed initial values $y_{0}$ and $\xi_{0}$, and with $u_{t}, \Phi_{t}$ and $\left[\eta_{t}^{\prime}, v_{t}^{\prime}\right]^{\prime}$ mutually independent. We define $\varepsilon_{t}:=\sum_{i=1}^{t} \eta_{i}$ with $\varepsilon_{0}=0_{p-r}$. The sequences $\varepsilon_{1: T}$ and $\xi_{1: T}$ are unobserved and take values $\varepsilon_{t} \in \mathcal{R}^{p-r}$ and $\xi_{t} \in \mathcal{R}^{r}$ for $0<r<p$. Additionally, the matrices are of dimensions $A \in \mathcal{R}^{p \times r}$ and $B \in \mathcal{R}^{p \times p-r}$, with $\left[\begin{array}{ll}A & B\end{array}\right] \in \mathcal{R}^{p \times p}$ and invertible. Let the random coefficient, $\Phi_{t}$, be i.i.d. Gaussian,

$$
\begin{equation*}
\operatorname{vec}\left(\Phi_{t}\right) \sim N\left(\operatorname{vec}(\Phi), \Omega_{\Phi}\right) \tag{3}
\end{equation*}
$$

with $\Omega_{\Phi}$ a positive definite covariance matrix. Let the observation error be i.i.d. Gaussian, such that $u_{t} \sim N\left(0, \Omega_{u}\right)$ with $\Omega_{u}$ a positive definite matrix, and let the innovations $\eta_{t}$ and $v_{t}$ be jointly Gaussian such that $\eta_{t} \sim N\left(0, \Omega_{\eta}\right)$ and $v_{t} \sim N\left(0, \Omega_{v}\right)$ with cross-covariance $\mathbb{C o v}\left[\eta_{t}, v_{t}\right]=\Omega_{\eta, v}$, such that the joint covariance matrix,

$$
\Lambda:=\left[\begin{array}{cc}
\Omega_{\eta} & \Omega_{\eta, v}  \tag{4}\\
\Omega_{\eta, v}^{\prime} & \Omega_{v}
\end{array}\right]
$$

is positive definite. Let all the introduced matrices be of appropriate dimensions and full rank. Furthermore, we introduce the orthogonal complements to $A$ and $B$, which we denote $b \in \mathcal{R}^{p \times r}$ and $a \in \mathcal{R}^{p \times p-r}$, such that $b^{\prime} B=0$ and $a^{\prime} A=0$ with $b$ and $a$ of full column rank. Finally, we let $C\left(y_{0}\right):=B\left(a^{\prime} B\right)^{-1} a^{\prime} y_{0}$.

Define the parameter vectors,

$$
\begin{gather*}
\omega:=\left[\begin{array}{lll}
\operatorname{vec}(B)^{\prime} & \operatorname{vec}(A)^{\prime} & \operatorname{vech}\left(\Omega_{u}\right)^{\prime}
\end{array}\right]^{\prime}  \tag{5}\\
\lambda:=\left[\begin{array}{llll}
\mu^{\prime} & \operatorname{vec}(\Phi)^{\prime} & \operatorname{vech}\left(\Omega_{\Phi}\right)^{\prime} & \operatorname{vech}(\Lambda)^{\prime}
\end{array}\right]^{\prime} \tag{6}
\end{gather*}
$$

which contain the parameters governing the observations $y_{t}$, and unobserved components $\varepsilon_{t}$ and $\xi_{t}$, respectively. The parameter vectors take values in $\omega \in \Theta_{\omega}$ and $\lambda \in \Theta_{\lambda}$, respectively. Additionally, we define the full parameter vector as

$$
\theta:=\left[\begin{array}{ll}
\omega^{\prime} & \lambda^{\prime} \tag{7}
\end{array}\right]^{\prime} \in \Theta_{\omega} \times \Theta_{\lambda}=: \Theta
$$

which indexes the model, and we refer to $\Theta$ as the parameter space. Note that $\omega$ and $\lambda$ in $\theta$ are variation free in the sense of Engle et al. (1983). The parameter space is a subset of the $d_{\theta}$-dimensional

Euclidean space $\Theta \subseteq \mathcal{R}^{d_{\theta}}$, where $d_{\theta}$ denotes the number of elements in $\theta$. In the case where no restrictions are imposed on $\theta$, the dimension $d_{\theta}$ increases rapidly in $r$ due to the $\frac{1}{2}\left(r^{2}+1\right) r^{2}$ parameters in $\Omega_{\Phi}$. We suggest restricting the off-diagonal elements of $\Omega_{\Phi}$ to zero to avoid over-parameterization. The number of parameters is then $d_{\theta}=2 p^{2}+p+2 r^{2}+r$ when the model is otherwise unrestricted.

The log-likelihood function for any parameter vector $\theta \in \Theta$, fixed initial values $y_{0} \in \mathcal{R}^{p}, \varepsilon_{0}=0_{p-r}$ and $\xi_{0} \in \mathcal{R}^{r}$, and observation sequence $y_{1: T} \in \mathcal{R}^{p \times T}$ is given by,

$$
\begin{equation*}
\boldsymbol{\ell}_{T}(\theta):=\log \boldsymbol{p}_{\theta}\left(\varepsilon_{0}, \xi_{0}, y_{0: T}\right) . \tag{8}
\end{equation*}
$$

The sample score is given by the first derivative of (8),

$$
\begin{equation*}
\boldsymbol{S}_{T}(\theta):=\frac{\partial}{\partial \theta} \ell_{T}(\theta), \tag{9}
\end{equation*}
$$

and the observed Information matrix is given by minus the second derivative of (8),

$$
\begin{equation*}
\boldsymbol{I}_{T}(\theta):=-\frac{\partial^{2}}{\partial \theta \partial \theta^{\prime}} \ell_{T}(\theta) \tag{10}
\end{equation*}
$$

Due to the nonlinear dynamics of the unobserved process (2), the log-likelihood function (8) and its derivatives (9)-(10) do not have closed-form solutions. In the following, we suppress the dependence on the initial values $\varepsilon_{0}, \xi_{0}$ and $y_{0}$, but note they remain fixed.

## 3. Properties of the Process

In this section we consider some properties of the process defined by Equations (1)-(3) for a given parameter value $\theta \in \Theta$. Specifically, we study the nonstationary and stationary components, including conditions on the parameter $\theta$ that ensure strict stationarity of the stationary component. Additionally, we decompose the observation $y_{t}$ into nonstationary and stationary directions.

### 3.1. The Unobserved Components

The first component of the model, $\varepsilon_{t}$, is a random walk (RW) in $p-r$ dimensions, equivalently expressed as an autoregressive process with a unit root. That is, for $t=1, \ldots, T$,

$$
\begin{equation*}
\varepsilon_{t}=\varepsilon_{t-1}+\eta_{t} \tag{11}
\end{equation*}
$$

with $\varepsilon_{0}=0_{p-r}$. The process (11) admits the transition density $p_{\lambda}\left(\varepsilon_{t} \mid \varepsilon_{t-1}\right)$ with respect to the $p-r$-dimensional Lebesgue measure; however, it does not have a stationary distribution. This type of process has been studied extensively, see e.g., Dickey and Fuller (1979). In summary, the RW process is linear and Gaussian, but nonstationary.

The second unobserved component of the model, $\xi_{t}$, is a random coefficient autoregressive (RCAR) process of lag order one in $r$ dimensions. The RCAR process (2)-(3) is observationally equivalent to a double autoregressive (DAR) process with one lag, cf. Ling (2007), which we formalize in Lemma 1.

Lemma 1. For $\theta \in \Theta$, the random coefficient autoregressive process (2)-(3) with $k=1$ has the following double autoregressive process representation, $t=1,2, \ldots, T$

$$
\begin{align*}
\xi_{t} & =\mu+\Phi \xi_{t-1}+\Omega_{v, t}^{1 / 2} z_{t}  \tag{12}\\
\Omega_{v, t} & =\Omega_{v}+\left(\xi_{t-1}^{\prime} \otimes I_{r}\right) \Omega_{\Phi}\left(\xi_{t-1}^{\prime} \otimes I_{r}\right)^{\prime} \tag{13}
\end{align*}
$$

for $\xi_{0}$ fixed, $z_{t} \sim N\left(0, I_{r}\right)$, cross-covariance $\operatorname{Cov}\left[\eta_{t}, z_{t}\right]=\Omega_{\eta v}$, and with the joint innovation process $\left[\eta_{t}^{\prime}, z_{t}^{\prime}\right]^{\prime}$ independent and identically distributed.

The DAR representation in Lemma 1 of the RCAR process in (2)-(3) characterizes the process dynamics in terms of the conditional mean and variance. The conditional mean $\mathbb{E}_{\lambda}\left[\xi_{t} \mid \xi_{t-1}\right]$ is autoregressive. However, the conditional variance $\operatorname{Var}_{\lambda}\left[\xi_{t} \mid \xi_{t-1}\right]$ depends positively on the lagged level 'squared'. The conditional variance is heteroskedastic, but not in the well-known ARCH sense of e.g., Engle (1982); rather, the lagged level of the process $\xi_{t-1}$ enters the variance, not the lagged innovation $v_{t-1}$. To illustrate the point, we consider for a moment the conditional variance in the univariate case $r=1$, which is given by $\omega_{v, t}^{2}=\omega_{v}^{2}+\omega_{\phi}^{2} \tilde{\xi}_{t-1}^{2}$. Here we see that a relatively large (in absolute terms) lagged level $\left|\xi_{t-1}\right|$ will result in a relatively large volatility $\omega_{v, t}$ in the present period, and vice versa.

We make the following assumption on the random coefficients (3) in order to ensure strict stationarity of the RCAR process (2)-(3).

Assumption 1. Assume that the top Lyapunov exponent is strictly negative,

$$
\begin{equation*}
\gamma:=\lim _{n \rightarrow \infty} \frac{1}{n} \mathbb{E}_{\lambda}\left[\log \left\|\prod_{t=1}^{n} \Phi_{t}\right\|\right]<0 \tag{14}
\end{equation*}
$$

Remark 1. The top Lyapunov exponent (14) is intractable but can be approximated to arbitrary precision via simulation, cf. inter alia Ling (2007) and Francq and Zakoian (2010). The following approximation converges almost surely

$$
\begin{equation*}
\hat{\gamma}_{n}:=\frac{1}{n} \log \left\|\prod_{t=1}^{n} \Phi_{t}\right\| \quad \xrightarrow{\text { a.s. }} \quad \gamma, \tag{15}
\end{equation*}
$$

as $n \rightarrow \infty$. In turn, $\hat{\gamma}_{n}$ can be computed efficiently via the QR-decomposition, cf. Dieci and Van Vleck (1995).
Assumption 1 ensures that the RCAR process can be characterized as a geometrically ergodic Markov chain, cf. Meyn and Tweedie (2005). This is formalized in the following theorem.

Theorem 1 (Feigin and Tweedie (1985), Theorem 3). Under Assumption 1, the process $\left\{\xi_{t}\right\}_{t=0,1, \ldots}$ is geometrically ergodic. In particular, the initial value $\tilde{\xi}_{0}$ can be given an initial distribution $p_{\theta}\left(\xi_{0}\right)$ such that $\left\{\xi_{t}\right\}_{t=0,1, \ldots}$ is stationary and geometrically ergodic with some fractional moment.

Remark 2. The stationary component, $\xi_{t}$, exhibits heavy-tailed behavior since it satisfies a stochastic recurrence equation. Pedersen and Wintenberger (2018) have recently considered the tail properties of processes of the form (2) for a more general specification of the random coefficient, $\Phi_{t}$, that includes BEKK-ARCH and DAR-type processes as special cases. It should be possible to show that the stationary distribution of $\xi_{t}$ as defined in (2)-(3) also has power-law tails under suitable conditions.

The RCAR process (2)-(3) admits the transition density $p_{\lambda}\left(\xi_{t} \mid \xi_{t-1}\right)$ with respect to the $r$-dimensional Lebesgue measure. Moreover, the process has the stationary distribution $\boldsymbol{p}_{\theta}\left(\xi_{t}\right)$ under Assumption 1. In summary, the RCAR process is Gaussian and strictly stationary, but nonlinear.

### 3.2. The Observed Process

The observations $\left\{y_{t}\right\}_{t=1,2, \ldots}$ are conditionally independent given the sequence of unobserved components $\left\{\varepsilon_{t}, \xi_{t}\right\}_{t=1,2, \ldots}$. Thus, the dynamics of the observed process are determined by the dynamics of the unobserved components.

We use the orthogonal complements $b^{\prime}$ and $a^{\prime}$ of the loading matrices $B$ and $A$, respectively, and the skew-projection identity of Johansen (1996) to decompose the observation vector $y_{t}$ as follows,

$$
\begin{equation*}
y_{t}=B_{a} a^{\prime} y_{t}+A_{b} b^{\prime} y_{t} \tag{16}
\end{equation*}
$$

where we define $B_{a}:=B\left(a^{\prime} B\right)^{-1}$ and $A_{b}:=A\left(b^{\prime} A\right)^{-1}$. Here $a^{\prime} B$ and $b^{\prime} A$ are invertible thanks to our assumption that $[A B]$ is square and invertible. By premultiplying $y_{t}$ by $a^{\prime}$ we eliminate the stationary directions, while leaving the nonstationary directions,

$$
\begin{equation*}
a^{\prime} y_{t}=a^{\prime} C\left(y_{0}\right)+a^{\prime} B \varepsilon_{t}+a^{\prime} u_{t} \tag{17}
\end{equation*}
$$

What is left after the linear transformation (17) is a random walk with Gaussian measurement error. Similarly, premultiplying $y_{t}$ by $b^{\prime}$ eliminates the nonstationary directions while the stationary directions remain,

$$
\begin{equation*}
b^{\prime} y_{t}=b^{\prime} A \xi_{t}+b^{\prime} u_{t} . \tag{18}
\end{equation*}
$$

The process given by (18) is a stationary random coefficient autoregressive process with Gaussian measurement error.

The decomposition of the observation process (16) allows for a cointegration interpretation of the SSR model. The $p$ observed variables in $y_{t}$ share $p-r$ common stochastic trends (17) with loading matrix $B_{a}$, while the $r$ linear combinations (18) are stationary and load into the levels with the matrix $A_{b}$. The observed process admits the conditional density $\boldsymbol{p}_{\theta}\left(y_{t} \mid y_{1: t-1}\right)$ with respect to the $p$-dimensional Lebesgue measure; however, this density does not have a closed-form expression. Moreover, the observed process does not have a stationary distribution.

## 4. Likelihood-Based Estimation and Inference

In this section, we introduce the ML estimator and consider its asymptotic properties. We wish to conduct estimation and inference based on the true, but intractable, model likelihood. Due to the intractability of the likelihood, we can neither compute the ML estimator via numerical optimization of (8), nor compute classic standard errors via the observed Information matrix (10). We refer to the ML estimator as being 'doubly intractable', with reference to the concept from the literature in Bayesian statistics on models with intractable likelihoods, see e.g., Murray et al. (2006). It is beyond the scope of this paper to derive a full asymptotic theory for the SSR model. Instead, we conjecture the limiting properties of the likelihood function (8) and its derivatives (9)-(10). We obtain the asymptotic properties for the ML estimator based on the conjecture.

We recall preliminarily that the ML estimator is defined as the parameter vector $\theta \in \Theta$ that maximizes the log-likelihood function (8),

$$
\begin{equation*}
\hat{\boldsymbol{\theta}}_{T}:=\underset{\theta \in \Theta}{\arg \sup } \boldsymbol{\ell}_{T}(\theta) \tag{19}
\end{equation*}
$$

noting that the ML estimator (19) is a function of the observation sequence $y_{1: T}$. We denote by $\theta^{*} \in \Theta$ the true parameter value for the data generating process (1)-(3). In the following, we make the below conjecture on the asymptotic properties of (8)-(10). Note that, having assumed that $B^{*}$ is known, the score, information, and likelihood in the conjecture refer to the unknown parameters only; that is, all elements in $\theta$ excluding vec $(B)$.

Conjecture 1. If Assumption 1 holds, $B^{*}$ is known, and $\theta^{*} \in \Theta \subseteq \mathcal{R}^{d_{\theta}}$, then the log-likelihood function $\ell_{T}(\cdot): \mathcal{R}^{d_{\theta}} \mapsto \mathcal{R}$ is three times continuously differentiable in $\theta$, and

1. $\frac{1}{\sqrt{T}} \boldsymbol{S}_{T}\left(\theta^{*}\right) \xrightarrow{D} N\left(0, \Omega_{S}\right)$ as $T \rightarrow \infty$, with $\Omega_{S}>0$,
2. $\frac{1}{T} \boldsymbol{I}_{T}\left(\theta^{*}\right) \xrightarrow{P} \Omega_{I}$ as $T \rightarrow \infty$, with $\Omega_{I}>0$, and
3. $\max _{h, i, j=1, \ldots, d_{\theta}} \sup _{\theta \in \mathcal{N}\left(\theta^{*}\right)}\left|\partial^{3} \ell_{T}(\theta) / \partial \theta_{h} \partial \theta_{i} \partial \theta_{j}\right| \leq c_{T}$,
where $\mathcal{N}\left(\theta^{*}\right)$ is a neighborhood of $\theta^{*}$ and $0 \leq c_{T} \xrightarrow{P} c, 0<c<\infty$, as $T \rightarrow \infty$.

Remark 3. Theorem 3 in Bohn Nielsen and Rahbek (2014) shows that Conjecture 1 holds in the case of the strictly stationary bivariate double autoregressive model with BEKK-type time-varying covariance. With B* known, the SSR model corresponds closely to this model plus Gaussian measurement errors.

It should be noted that we propose Conjecture 1 despite lack of finite moments of the RCAR process, cf. Theorem 1. This is in line with the results of inter alia Bohn Nielsen and Rahbek (2014) for the bivariate DAR model, and Ling $(2004,2007)$ for the univariate DAR model.

The result in Theorem 2 below states that if Conjecture 1 holds true, then the ML estimator (19) is unique, $\sqrt{T}$-consistent and asymptotically Gaussian. The result follows from applying Lemma 1 in Jensen and Rahbek (2004), the conditions of which correspond to (1.)-(3.) of Conjecture 1.

Theorem 2 (Jensen and Rahbek (2004), Lemma 1). If Conjecture 1 holds, then there exists a fixed open neighborhood $\mathcal{U}\left(\theta^{*}\right) \subseteq \mathcal{N}\left(\theta^{*}\right)$ of the true parameter $\theta^{*}$, which is an interior point of $\Theta$, such that with probability tending to one as $T \rightarrow \infty$, there exists a minimum point $\hat{\boldsymbol{\theta}}_{T}$ in $\mathcal{U}\left(\theta^{*}\right)$ and $\boldsymbol{\ell}_{T}(\theta)$ is convex in $\mathcal{U}\left(\theta^{*}\right)$. In particular, $\hat{\boldsymbol{\theta}}_{T}$ is unique and satisfies the score equation

$$
\begin{equation*}
\boldsymbol{S}_{T}\left(\hat{\boldsymbol{\theta}}_{T}\right)=0 \tag{20}
\end{equation*}
$$

Additionally, the ML estimator is consistent $\hat{\boldsymbol{\theta}}_{T} \rightarrow \theta^{*}$, and asymptotically Gaussian,

$$
\begin{equation*}
\sqrt{T}\left(\hat{\boldsymbol{\theta}}_{T}-\theta^{*}\right) \quad \xrightarrow{D} \quad N\left(0, \Omega_{I}^{-1} \Omega_{S} \Omega_{I}^{-1}\right), \quad T \rightarrow \infty . \tag{21}
\end{equation*}
$$

Proof. Conjecture 1 satisfies the Cramer-type conditions of Lemma 1 in Jensen and Rahbek (2004), which provides the result.

We assume that the true value of $B$ is known, because Chang et al. (2009) showed that the ML estimator of the loading matrix $B$ exhibits $T$-convergence and is asymptotically mixed Gaussian in the CST model. The CST model corresponds to the SSR model with $p-r=1$, but without the stationary components, i.e., $A=0_{p \times r}$ for any $p$. We find it reasonable to believe that this result carries over to the SSR model. Moreover, fixing $B$ is conceptually similar to classic cointegration analysis with known cointegrating vectors, which is an accepted starting point for new methodological developments, see e.g., Bec and Rahbek (2004). In applications we often have a predefined set of cointegrating vectors that we are interested in. In the context of the SSR model, the cointegrating vectors correspond to the rows of the orthogonal complement $b^{\prime}$. As an example, for the empirical illustration in Section 10 we consider an interest rate spread in a bivariate system with one common stochastic trend, i.e., $p=2$ and $p-r=1$. The spread implies $b^{\prime}=\left[\begin{array}{ll}1 & -1\end{array}\right]$, which in turn corresponds to the loading matrix $B=\left[\begin{array}{ll}1 & 1\end{array}\right]^{\prime}$ when normalizing on the first element.

The Fisher Information matrix, $\Omega_{I}$, is consistently estimated by the (scaled) observed Information matrix evaluated at $\hat{\boldsymbol{\theta}}_{T}$, cf. Conjecture 1.(3.). Moreover, the asymptotic variance of the score, $\Omega_{S}$, is equal to the Fisher Information matrix when the model is well-specified; the information matrix equality holds, cf. e.g., Hamilton (1994, sct. 14.4). In this case, the asymptotic variance of the ML estimator (19) is simply the inverse Fisher Information matrix. Thus, we can use classic standard errors, that are based on the observed Information matrix (10), to conduct inference on the ML estimates.

## 5. The Incomplete Data Framework

In this section, we appeal to the incomplete data framework of Dempster et al. (1977) to deal with the unobserved components of the SSR model. We first formulate the state space representation of the model in (1)-(3) and its associated optimal filtering problem. Secondly, we formulate the intractable sample score (9) and observed information matrix (10) in terms of the optimal filtering problem. In Section 6 we introduce a particle filter algorithm with which we can approximate the
optimal filtering problem. This enables approximation of the intractable sample score and observed information matrix via the particle filter algorithm.

### 5.1. The State Space Form and the Optimal Filtering Problem

Preliminarily, we collect the unobserved components in the vector $x_{t}:=\left[\begin{array}{cc}\varepsilon_{t}^{\prime} & \xi_{t}^{\prime}\end{array}\right]^{\prime}$, which we refer to as the state vector. The unobserved components are Markov, see (11)-(13), and the observation depends only on the contemporary values of the unobserved components. Thus, the SSR model in (1)-(3) has the dependency structure of a state space model. Formally, for $t=1, \ldots, T$, the SSR model in (1)-(3) has the following state space representation,

$$
\begin{gather*}
y_{t}=C\left(y_{0}\right)+\Pi x_{t}+\Omega_{u}^{1 / 2} u_{t}  \tag{22}\\
x_{t}=\alpha+\Gamma x_{t-1}+\Lambda_{t}^{1 / 2} v_{t} \tag{23}
\end{gather*}
$$

with $y_{0}$ and $x_{0}$ fixed, $u_{t} \sim N\left(0, I_{p}\right)$ and $v_{t} \sim N\left(0, I_{p}\right)$, and $u_{t}$ and $v_{t}$ mutually independent. We define accordingly,

$$
\Pi:=\left[\begin{array}{l}
B^{\prime}  \tag{24}\\
A^{\prime}
\end{array}\right]^{\prime}, \quad \alpha:=\left[\begin{array}{l}
0 \\
\mu
\end{array}\right], \quad \Gamma:=\left[\begin{array}{cc}
I_{p-r} & 0 \\
0 & \Phi
\end{array}\right] \quad \text { and } \quad \Lambda_{t}:=\left[\begin{array}{cc}
\Omega_{\eta} & \Omega_{\eta, v} \\
\Omega_{\eta, v}^{\prime} & \Omega_{v, t}
\end{array}\right]
$$

and recall that $\Omega_{v, t}$ is defined in Lemma (1). We refer to (22) as the observation equation, and to (23) as the transition equation. It is easy to verify that the state space representation in (22) and (23) is observationally equivalent to the SSR model as presented in (1)-(3). The observation and transition equations admit the densities with respect to the $p$-dimensional Lebesgue measure,

$$
\begin{align*}
& p_{\omega}\left(y_{t} \mid x_{t}\right)=N\left(y_{t} ; C\left(y_{0}\right)+\Pi x_{t}, \Omega_{u}\right)  \tag{25}\\
& p_{\lambda}\left(x_{t} \mid x_{t-1}\right)=N\left(x_{t} ; \alpha+\Gamma x_{t-1}, \Lambda_{t}\right) \tag{26}
\end{align*}
$$

respectively. We refer to (25) as the observation density and to (26) as the transition density. As mentioned previously, we suppress the dependence on the initial observation $y_{0}$.

One approach to conducting inference on the unobserved components, i.e., the state vector $x_{t}$, is the optimal filtering problem, cf. Anderson and Moore (1979). The optimal filtering problem refers to the general problem of computing the conditional expectation of some sequence of unobserved states given some sequence of observations. In the following, we consider the specific instance of the optimal filtering problem known as the smoothing problem. Formally, the smoothing problem is a conditional expectation of the form,

$$
\begin{equation*}
\mathbb{E}_{\theta}\left[\gamma_{t}\left(x_{1: t}\right) \mid y_{1: t}\right]=\int \gamma_{t}\left(x_{1: t}\right) \boldsymbol{p}_{\theta}\left(x_{1: t} \mid y_{1: t}\right) \mathrm{d} x_{1: t} \tag{27}
\end{equation*}
$$

for any function $\gamma_{t}\left(x_{1: t}\right) \in L^{1}\left[\mathcal{R}^{t p}, \boldsymbol{p}_{\theta}\left(x_{1: t} \mid y_{1: t}\right)\right]$ and point in time $t \in\{1, \ldots, T\}$. We refer to the function $\gamma_{t}\left(x_{1: t}\right)$ as the test function and to the density $\boldsymbol{p}_{\theta}\left(x_{1: t} \mid y_{1: t}\right)$ as the smoothing density. The test function may be time-varying, but of known form for a fixed observation sequence $y_{1: T}$. The smoothing density in (27) can be expressed as the recursion of the lagged smoothing density,

$$
\begin{equation*}
\boldsymbol{p}_{\theta}\left(x_{1: t} \mid y_{1: t}\right)=\frac{p_{\omega}\left(y_{t} \mid x_{t}\right) p_{\lambda}\left(x_{t} \mid x_{t-1}\right)}{\boldsymbol{p}_{\theta}\left(y_{t} \mid y_{1: t-1}\right)} \boldsymbol{p}_{\theta}\left(x_{1: t-1} \mid y_{1: t-1}\right) \tag{28}
\end{equation*}
$$

initialized with $\boldsymbol{p}_{\theta}\left(x_{1} \mid x_{0}, y_{0}, y_{1}\right)$. The normalizing constant in (28) is the likelihood contribution, which is given by the integral,

$$
\begin{equation*}
\boldsymbol{p}_{\theta}\left(y_{t} \mid y_{1: t-1}\right)=\int p_{\omega}\left(y_{t} \mid x_{t}\right) p_{\lambda}\left(x_{t} \mid x_{t-1}\right) \boldsymbol{p}_{\theta}\left(x_{1: t-1} \mid y_{1: t-1}\right) \mathrm{d} x_{1: t} \tag{29}
\end{equation*}
$$

We note the smoothing density recursion (28) is intractable due to the intractability of the likelihood contribution (29). In the following, we will use the smoothing problem (27) to address computation of the sample score (9) and observed Information matrix (10).

### 5.2. The Sample Score and Observed Information as Smoothing Problems

The incomplete data framework is closely associated with the classic expectation maximization (EM) algorithm, introduced in Dempster et al. (1977). The EM algorithm is a common approach to maximizing the log-likelihood function (8) to obtain the ML estimator (19) for models with unobserved variables. When the EM algorithm is applicable, it is also possible to evaluate the sample score (9) and observed Information matrix (10). For the SSR model, however, the EM algorithm does not apply directly, yet we may use the incomplete data framework to reformulate the sample score and observed Information in terms of intractable smoothing problems of the form (27).

A central concept of the EM algorithm is the auxiliary function called the intermediate quantity, which is defined as,

$$
\begin{align*}
Q_{T}(\theta \mid \vartheta) & :=\int \log p_{\theta}\left(y_{1: T}, x_{1: T}\right) \boldsymbol{p}_{\vartheta}\left(x_{1: T} \mid y_{1: T}\right) \mathrm{d} x_{1: T} \\
& =\ell_{T}(\theta)-\boldsymbol{H}_{T}(\theta \mid \vartheta) \tag{30}
\end{align*}
$$

where

$$
\begin{equation*}
\boldsymbol{H}_{T}(\theta \mid \vartheta):=-\int \log p_{\theta}\left(x_{1: T} \mid y_{1: T}\right) \boldsymbol{p}_{\vartheta}\left(x_{1: T} \mid y_{1: T}\right) \mathrm{d} x_{1: T} \tag{31}
\end{equation*}
$$

for any parameter values $\theta, \vartheta \in \Theta$. We refer to $\log p_{\theta}\left(y_{1: T}, x_{1: T}\right)$ as the complete data log-likelihood. By the state space model structure (22)-(23) and variation freeness of $\theta$ defined in (7), we have that the complete data log-likelihood is given by,

$$
\begin{equation*}
\log p_{\theta}\left(y_{1: T}, x_{1: T}\right)=\sum_{t=1}^{T}\left[\log p_{\omega}\left(y_{t} \mid x_{t}\right)+\log p_{\lambda}\left(x_{t} \mid x_{t-1}\right)\right] \tag{32}
\end{equation*}
$$

The intermediate quantity (30) is sometimes also called the expected log-likelihood, since it is interpretable as the conditional expectation of the complete data log-likelihood (32) given the observations $y_{1: T}$. We note the term separating the log-likelihood (8) and the intermediate quantity (30) is the entropy of the smoothing density (28) with parameters $\vartheta$ and $\theta$, defined in (31).

We are interested in the intermediate quantity (30) because it provides a convenient way to derive the sample score and observed Information matrix in terms of the derivatives of the complete data $\log$-likelihood (32). The first and second derivatives of the complete data log-likelihood function in (32) are the sum of the first and second order derivatives of the observation and transition log-densities with respect to $\omega$ and $\lambda$, respectively. These can be computed by either analytical or numerical differentiation of (32). For $\vartheta \in \Theta$, we define the derivatives of (32) in terms of the functions,

$$
\begin{align*}
& U_{T}\left(x_{1: T} ; \vartheta\right):=\left.\frac{\partial}{\partial \theta} \log p_{\theta}\left(y_{1: T}, x_{1: T}\right)\right|_{\theta=\vartheta}=\sum_{t=1}^{T} u_{t}\left(x_{t}, x_{t-1} ; \vartheta\right)  \tag{33}\\
& V_{T}\left(x_{1: T} ; \vartheta\right):=\left.\frac{\partial^{2}}{\partial \theta \partial \theta^{\prime}} \log p_{\theta}\left(y_{1: T}, x_{1: T}\right)\right|_{\theta=\vartheta}=\sum_{t=1}^{T} v_{t}\left(x_{t}, x_{t-1} ; \vartheta\right) \tag{34}
\end{align*}
$$

where, taking advantage of the variation freeness of the model parameter, $\theta$, we define the summands of (33) and (34), respectively, as

$$
u_{t}\left(x_{t}, x_{t-1} ; \vartheta\right):=\left.\left[\begin{array}{c}
\frac{\partial}{\partial \omega} \log p_{\omega}\left(y_{t} \mid x_{t}\right)  \tag{35}\\
\frac{\partial}{\partial \lambda} \log p_{\lambda}\left(x_{t} \mid x_{t-1}\right)
\end{array}\right]\right|_{\theta=\vartheta}
$$

and

$$
v_{t}\left(x_{t}, x_{t-1} ; \vartheta\right):=\left.\left[\begin{array}{cc}
\frac{\partial^{2}}{\partial \omega \partial \omega^{\prime}} \log p_{\omega}\left(y_{t} \mid x_{t}\right) & 0_{d_{\omega} \times d_{\lambda}}  \tag{36}\\
0_{d_{\lambda} \times d_{\omega}} & \frac{\partial^{2}}{\partial \lambda \partial \lambda^{\prime}} \log p_{\lambda}\left(x_{t} \mid x_{t-1}\right)
\end{array}\right]\right|_{\theta=\vartheta}
$$

We note that the functions (35) and (36) should not be confused with the measurement error in (22) and innovations in (23), respectively.

If the first and second order derivatives of the complete data log-likelihood in (33) and (34), respectively, are integrable with respect to the smoothing density (28), then we may appeal to Fisher's and Louis' identities (defined below) to express the sample score (9) and observed Information matrix (10) in terms of smoothing problems of the form (27).

Conjecture 2. For any $\theta \in \Theta$ and observation sequence $y_{1: T} \in \mathcal{R}^{p \times T}$, it holds that $U_{T}\left(x_{1: T} ; \theta\right) \in$ $L^{2}\left[\mathcal{R}^{p \times T}, \boldsymbol{p}_{\theta}\left(x_{1: T} \mid y_{1: T}\right)\right]$ and $V_{T}\left(x_{1: T} ; \theta\right) \in L^{1}\left[\mathcal{R}^{p \times T}, \boldsymbol{p}_{\theta}\left(x_{1: T} \mid y_{1: T}\right)\right]$.

For the same reasons we conjectured the asymptotic properties of the true log-likelihood function, sample score, observed information matrix, we conjecture integrability of the derivatives of the complete data log-likelihood (33) and (34).

Fisher's identity, cf. Dempster et al. (1977), states the first derivative of the intermediate quantity (30) is equivalent to the sample score (9). Similarly, Louis' identity of Louis (1982) establishes a relation between the first and second derivatives of the intermediate quantity (30) and the observed Information matrix (10).

Lemma 2 (Fisher's and Louis' identities, cf. Cappé et al. (2005), Proposition 10.1.6). If Conjecture 2 holds and $\theta \in \Theta$, then the sample score (9) is equivalently given by

$$
\begin{equation*}
\boldsymbol{S}_{T}(\theta)=\int U_{T}\left(x_{1: T} ; \theta\right) \boldsymbol{p}_{\theta}\left(x_{1: T} \mid y_{1: T}\right) \mathrm{d} x_{1: T} \tag{37}
\end{equation*}
$$

and the observed Information (10) is equivalently given by

$$
\begin{equation*}
\boldsymbol{I}_{T}(\theta)=\boldsymbol{S}_{T}(\theta) \boldsymbol{S}_{T}(\theta)^{\prime}-\boldsymbol{G}_{T}(\theta)-\boldsymbol{K}_{T}(\theta), \tag{38}
\end{equation*}
$$

where

$$
\begin{align*}
& \boldsymbol{G}_{T}(\theta):=\int V_{T}\left(x_{1: T} ; \theta\right) \boldsymbol{p}_{\theta}\left(x_{1: T} \mid y_{1: T}\right) \mathrm{d} x_{1: T}  \tag{39}\\
& \boldsymbol{K}_{T}(\theta):=\int U_{T}\left(x_{1: T} ; \theta\right) U_{T}\left(x_{1: T} ; \theta\right)^{\prime} \boldsymbol{p}_{\theta}\left(x_{1: T} \mid y_{1: T}\right) \mathrm{d} x_{1: T} \tag{40}
\end{align*}
$$

and the functions $U_{T}\left(x_{1: T} ; \theta\right)$ and $V_{T}\left(x_{1: T} ; \theta\right)$ are defined in (33) and (34), respectively.
Although Lemma 2 shows the sample score (9) and observed Information (10) can be restated as smoothing problems of the form (27), we still cannot obtain closed-form expressions due to the intractability of the optimal filtering problem, cf. Section 5.1. In the next section, we introduce a particle filter algorithm that can approximate smoothing problems for appropriately chosen test functions, such as the functions $U_{T}\left(x_{1: T} ; \theta\right)$ and $V_{T}\left(x_{1: T} ; \theta\right)$ under Conjecture 2.

## 6. Particle Filter-Based Approximations

In this section, we introduce a particle filter algorithm that produces pointwise approximations to the true but intractable log-likelihood function (8), sample score (9), and observed Information matrix (10) for any parameter $\theta \in \Theta$ and fixed observation sequence $y_{1: T} \in \mathcal{R}^{p \times T}$. In Section 7 , we show how to apply the particle filter-based approximations introduced in this section to approximate the true, intractable ML estimator and classic standard errors, which we introduced in Section 4.

### 6.1. Particle Filtering

A particle filter is a simulation-based algorithm that produces approximations to smoothing problems of the form (27) for state space models. We introduce here a standard particle filter, which produces empirical measures that recursively approximate the smoothing density (28) for each time point in the observed sample $t \in\{1, \ldots, T\}$. The empirical measures consist of point masses, which we refer to as particles, and we use these for Monte Carlo integration in order to approximate the smoothing problem (27). Additionally, the particle filter produces a point-wise approximation of the log-likelihood function as a by-product. For an introduction to particle filtering in the context of economics and finance see Creal (2012).

The particle filter algorithm relies on an importance density, denoted $q_{\theta}\left(x_{1: t} \mid y_{1: t}\right)$, that has the same support and recursive structure as the smoothing density (28). Formally, for $t=1, \ldots, T$, we define the importance density as,

$$
\begin{equation*}
q_{\theta}\left(x_{1: t} \mid y_{1: t}\right):=q_{\theta}\left(x_{t} \mid x_{t-1}, y_{t}\right) q_{\theta}\left(x_{1: t-1} \mid y_{1: t-1}\right) \tag{41}
\end{equation*}
$$

initialized by $q_{\theta}\left(x_{1} \mid x_{0}, y_{0}, y_{1}\right)$. We note the importance density (41) is defined recursively by $q_{\theta}\left(x_{t} \mid x_{t-1}, y_{t}\right)$, which we refer to as the importance transition density.

Assuming the smoothing density (28) is absolutely continuous with respect to the importance density (41), we can write the former as a the product of the importance density and a weight function,

$$
\begin{equation*}
\boldsymbol{p}_{\theta}\left(x_{1: t} \mid y_{1: t}\right)=\overline{\boldsymbol{w}}_{t}\left(x_{1: t}\right) q_{\theta}\left(x_{1: t} \mid y_{1: t}\right), \quad \overline{\boldsymbol{w}}_{t}\left(x_{1: t}\right):=\frac{\boldsymbol{p}_{\theta}\left(x_{1: t} \mid y_{1: t}\right)}{q_{\theta}\left(x_{1: t} \mid y_{1: t}\right)} \tag{42}
\end{equation*}
$$

We refer to the weight function $\bar{w}_{t}\left(x_{1: t}\right)$ as the normalized importance weight. We note that (42) constitutes a change of measure from the smoothing density to the importance density, and the normalized importance weight is a Radon-Nikodym derivative between the two densities.

Substituting the recursive expressions for the smoothing density (28) and importance density (41) into the expression for the normalized importance weight in (42), we obtain a recursive expression for the normalized importance weight,

$$
\begin{equation*}
\overline{\boldsymbol{w}}_{t}\left(x_{1: t}\right)=\frac{\tilde{w}_{t}\left(x_{t-1: t}\right)}{\boldsymbol{p}_{\theta}\left(y_{t} \mid y_{1: t-1}\right)} \overline{\boldsymbol{w}}_{t-1}\left(x_{1: t-1}\right), \tag{43}
\end{equation*}
$$

where we define

$$
\begin{equation*}
\tilde{w}_{t}\left(x_{t-1: t}\right):=\frac{p_{\omega}\left(y_{t} \mid x_{t}\right) p_{\lambda}\left(x_{t} \mid x_{t-1}\right)}{q_{\theta}\left(x_{t} \mid x_{t-1}, y_{t}\right)} \tag{44}
\end{equation*}
$$

We refer to (44) as the incremental importance weights. The recursion for the normalized importance weight (43) is normalized by the likelihood contribution (29) and is therefore also intractable.

For particle filtering in general, the importance transition density is subject to choice under mild regularity conditions, cf. e.g., Assumption 9.4.1 in Cappé et al. (2005). We let the importance transition density be the corresponding model density; formally,

$$
\begin{equation*}
q_{\theta}\left(x_{t} \mid x_{t-1}, y_{t}\right):=p_{\theta}\left(x_{t} \mid x_{t-1}, y_{t}\right) . \tag{45}
\end{equation*}
$$

We refer to (45) as the locally optimal transition density. This choice of importance transition density is optimal in the sense that it is conditional on the the contemporary observation $y_{t}$, cf. Doucet et al. (2000). This is sometimes also referred to as 'fully adapted', cf. e.g., Pitt and Shephard (1999b). If we instead let the importance transition density be the model transition density (26), we omit the information about $x_{t}$ that is contained in $y_{t}$. The locally optimal transition density is not necessarily available in closed-form for nonlinear state space models. It is, however, available for the SSR model and we present it in Lemma 3.

Lemma 3. For $\theta \in \Theta$, the locally optimal transition density has the closed-form expression

$$
\begin{equation*}
p_{\theta}\left(x_{t} \mid x_{t-1}, y_{t}\right)=N\left(x_{t} ; \mu_{t \mid t}^{x}, \Sigma_{t \mid t}^{x}\right), \tag{46}
\end{equation*}
$$

where the conditional mean and variance are given by,

$$
\begin{gather*}
\mu_{t \mid t}^{x}=\mu_{t \mid t-1}^{x}+\Sigma_{t \mid t-1}^{x} \Pi^{\prime}\left[\Sigma_{t \mid t-1}^{y}\right]^{-1}\left(y_{t}-\mu_{t \mid t-1}^{y}\right)  \tag{47}\\
\Sigma_{t \mid t}^{x}=\Sigma_{t \mid t-1}^{x}-\Sigma_{t \mid t-1}^{x} \Pi^{\prime}\left[\Sigma_{t \mid t-1}^{y}\right]^{-1} \Pi \Sigma_{t \mid t-1}^{x}, \tag{48}
\end{gather*}
$$

with

$$
\begin{gather*}
\mu_{t \mid t-1}^{y}=C\left(y_{0}\right)+\Pi \mu_{t \mid t-1}^{x}  \tag{49}\\
\Sigma_{t \mid t-1}^{y}=\Pi \Sigma_{t \mid t-1}^{x} \Pi^{\prime}+\Omega_{u}  \tag{50}\\
\mu_{t \mid t-1}^{x}=\alpha+\Gamma x_{t-1}  \tag{51}\\
\Sigma_{t \mid t-1}^{x}=\Lambda_{t}, \tag{52}
\end{gather*}
$$

and the state space form definitions given in (24).
Remark 4. The locally optimal transition density (46) is related to the Kalman (1960) filter, which solves the optimal filtering problem analytically for linear and Gausian models. Equations (49)-(52) correspond the Kalman filter for a known value of $x_{t-1}$. Related methods for efficient particle filtering include the mixture Kalman filter and Rao-Blackwellisation, cf. Chen and Liu (2000) and Andrieu and Doucet (2002).

It is straightforward to use the general expression for the incremental importance weight in (44) to show that letting the importance transition density be the locally optimal transition density, i.e., (45), results in the following specific expression for incremental importance weights,

$$
\begin{equation*}
\tilde{w}_{t}\left(x_{t-1}\right)=p_{\theta}\left(y_{t} \mid x_{t-1}\right) . \tag{53}
\end{equation*}
$$

We refer to the density in (53) as the predictive observation density. It has a closed-form expression that follows from the closed-form expression of the locally optimal transition density in Lemma 3.

Corollary 1. For $\theta \in \Theta$, the predictive observation density has the closed-form expression

$$
\begin{equation*}
p_{\theta}\left(y_{t} \mid x_{t-1}\right)=N\left(y_{t} ; \mu_{t \mid t-1}^{y}, \Sigma_{t \mid t-1}^{y}\right) \tag{54}
\end{equation*}
$$

recalling the definitions in (49)-(52).
Proof. Contained in the proof of Lemma 3.
Remark 5. The choice of importance transition density (45) is locally optimal in the sense that the conditional variance of the incremental importance weights (53) given $x_{t-1}$ is zero, cf. Doucet et al. (2000).

The particle filter, presented in Algorithm 1 below, produces weighted particle samples approximately distributed as the smoothing density (28) at each point in time $t=1, \ldots, T$. The algorithm consists of iterating over three steps. At point $t$ in time, the first step is to sample $N$ particles, denoted $\left\{\tilde{x}_{1: t}^{(i)}\right\}_{i=1}^{N}$, from the importance density (41) given the particle sample from $t-1$. This is called the propagation step. Step two consists of computing self-normalized importance weights, denoted $\left\{\bar{w}_{t}^{(i)}\right\}_{i=1}^{N}$, that approximate the normalized importance weights (43). This is the weighting step. The third step is to sample $N$ particle indices, denoted $\left\{I^{(i)}\right\}_{i=1}^{N}$, with replacement. We sample index $j$ with probability $\bar{w}_{t}^{(j)}$ for $j \in\{1, \ldots, N\}$. We retain the number of particles indicated by the resulting sample of particle indices, denoted $\left\{x_{1: t}^{(i)}\right\}_{i=1}^{N}$, and let the importance weights be uniform. This is the resampling step. After resampling, we store the particle samples and proceed to $t+1$.

For a fixed parameter value $\theta \in \Theta$ and observation sequence $y_{1: T} \in \mathcal{R}^{p \times T}$, we run the locally optimal particle filter for the SSR model as specified in Algorithm 1 below.

## Algorithm 1: Locally Optimal Particle Filter.

Given a parameter $\theta \in \Theta$, initialize by setting $x_{0}^{(i)}:=x_{0}$ and $\bar{w}_{0}^{(i)}:=1 / N$ for $i=1, \ldots, N$. For $t=0,1, \ldots, T$ :

1. Sample particles $\left\{\tilde{x}_{t}^{(i)}\right\}_{i=1}^{N}$ with distribution

$$
\begin{equation*}
\tilde{x}_{t}^{(i)} \sim p_{\theta}\left(x_{t} \mid x_{t-1}^{(i)}, y_{t}\right), \tag{55}
\end{equation*}
$$

and set $\tilde{x}_{1: t}^{(i)}:=\left[\begin{array}{cc}x_{1: t-1}^{(i)} & \tilde{x}_{t}^{(i)}\end{array}\right]$ for $i=1,2, \ldots, N$.
2. Calculate the unnormalized importance weights, $\left\{w_{t}^{(i)}\right\}_{i=1}^{N}$,

$$
\begin{equation*}
w_{t}^{(i)}=p_{\theta}\left(y_{t} \mid \tilde{x}_{t-1, t}^{(i)}\right) \bar{w}_{t-1}^{(i)}, \tag{56}
\end{equation*}
$$

for $i=1, \ldots, N$. Then compute the normalized importance weights

$$
\begin{equation*}
\bar{w}_{t}^{(i)}=\frac{w_{t}^{(i)}}{W_{t}^{N}}, \quad W_{t}^{N}:=\sum_{i=1}^{N} w_{t}^{(i)}, \tag{57}
\end{equation*}
$$

for $i=1, \ldots, N$.
3. Sample $N$ particle indices $\left\{I^{(i)}\right\}_{i=1}^{N}, I^{(i)} \in\{1, \ldots, N\}$, with probabilities

$$
\begin{equation*}
\operatorname{Pr}\left(I^{(i)}=j \mid \tilde{\mathcal{F}}_{t}, y_{1: t}\right)=\bar{w}_{t}^{(j)}, \quad j \in\{1, \ldots, N\} \tag{58}
\end{equation*}
$$

for $i=1, \ldots, N$. Set the resampled particles $x_{1: t}^{(i)}:=\tilde{x}_{1: t}^{\left(I^{(i)}\right)}$, and the normalized importance weights $\bar{w}_{t}^{(i)}:=1 / N$ for $i=1, \ldots, N$.

Remark 6. The resampling method applied in step (3.) of Algorithm 1 is known as multinomial resampling. Alternative methods that are guaranteed to produce lower Monte Carlo variance exists, cf. Douc et al. (2005). We consider multinomial resampling for its analytical tractability, and recommend applying one of the more efficient alternatives in practice.

Remark 7. The notation $x_{1: t}^{(i)}$ is ambiguous due to the resampling step of Algorithm 1 , since the elements of the ith particle path at time $t-1$, denoted $x_{1: t-1}^{(i)}$, are not necessarily the same as the first $t-1$ elements of the ith particle path at time $t$, denoted $x_{1: t}^{(i)}$. By convention, $x_{1: t}^{(i)}$ always refers to the particle chain after resampling at
time $t$ (similarly $\tilde{x}_{1: t}^{(i)}$ refers to the chain before resampling). We refer to elements $k$ to $l$ of the ith particle chain after resampling at time $t$ as $x_{l: k, t}^{(i)}$.

The particle filter in Algorithm 1 produces two particle samples at each point in time, $t$. The first set, $\left\{\tilde{x}_{1: t}^{(i)}\right\}_{i=1}^{N}$, is produced at the propagation step (1.) and is associated with importance weights in the weighting step (2.), $\left\{\bar{w}_{t}^{(i)}\right\}_{i=1}^{N}$. The second set, $\left\{x_{1: t}^{(i)}\right\}_{i=1}^{N}$, is produced at the resampling step (3.). Both sets are approximately drawn from the smoothing density (28). We note the resampling step introduces additional sampling error, cf. Chopin (2004), so we calculate approximations using the weighted sample unless otherwise specified.

The particle filter iterates over over the propagation, weighting and resampling steps throughout the sequence, $t=1, \ldots, T$, after which the algorithm terminates. We note the two sets of particles produced during each iteration are themselves random variables measurable with respect to the sub- $\sigma$-algebras $\tilde{\mathcal{F}}_{t}$ and $\mathcal{F}_{t}$, defined next.

Definition 1. Define the sub- $\sigma$-algebras $\tilde{\mathcal{F}}_{t}:=\mathcal{F}_{t-1} \cup \sigma\left(\tilde{x}_{t}^{(1)}, \ldots, \tilde{x}_{t}^{(N)}\right), \mathcal{F}_{t}:=\tilde{\mathcal{F}}_{t} \cup \sigma\left(x_{t}^{(1)}, \ldots, x_{t}^{(N)}\right)$ for $t=1, \ldots, T$, initialized by $\mathcal{F}_{0}:=\varnothing$.

At each point in time, we associate an empirical measure with the weighted particle sample generated by the propagation (1.) and reweighting (2.) steps in Algorithm 1. Formally, for $t=$ $1,2, \ldots, T$, we define the empirical measure,

$$
\begin{equation*}
\tilde{p}_{\theta}^{N}\left(\mathrm{~d} x_{1: t} \mid y_{1: t}\right):=\sum_{i=1}^{N} \bar{w}_{t}^{(i)} \delta_{\tilde{x}_{1: t}^{(i)}}\left(\mathrm{d} x_{1: t}\right), \tag{59}
\end{equation*}
$$

where $\delta_{x^{\prime}}(\mathrm{d} x)$ denotes the point measure at $x^{\prime} \in \mathcal{R}^{p}$ with respect to $\mathrm{d} x$. The weighted particles that constitute the empirical measure (59) are approximately distributed according to the smoothing density (28). We emphasize the weighted particles are not independent draws from (28), because the resampling step introduces dependence between the particles at each iteration of the algorithm. We use the empirical measure (59) to define a particle filter-based approximation of the intractable smoothing problem in (27),

$$
\begin{equation*}
\tilde{\mathbb{E}}_{\theta}^{N}\left[\gamma_{t}\left(x_{1: t}\right) \mid y_{1: t}\right]:=\int \gamma_{t}\left(x_{1: t}\right) \tilde{p}_{\theta}^{N}\left(\mathrm{~d} x_{1: t} \mid y_{1: t}\right)=\sum_{i=1}^{N} \bar{w}_{t}^{(i)} \gamma_{t}\left(\tilde{x}_{1: t}^{(i)}\right), \tag{60}
\end{equation*}
$$

for any point in time $t \in\{1, \ldots, T\}$. Due to dependence between the weighted particles, we cannot establish the asymptotic properties of the approximation (60) based on the law of large numbers and central limit theorem for independent random variables. For appropriately chosen test functions $\gamma_{t}\left(x_{1: t}\right)$, the approximation (60) is both consistent and asymptotically Gaussian as the number of particles tends to infinity, $N \rightarrow \infty$, cf. Theorem 9.4.5 in Cappé et al. (2005).

The particle filter in Algorithm 1 also produces an approximation of the log-likelihood function (8) evaluated at the parameter value $\theta$ and the observation sequence $y_{1: T}$,

$$
\begin{equation*}
\tilde{\ell}_{T}^{N}(\theta):=\sum_{t=1}^{T} \log W_{t}^{N} \tag{61}
\end{equation*}
$$

We note that the approximate log-likelihood function (61) consists of the logarithm of the product of normalizing constants produced by Algorithm 1. The approximate log-likelihood (61) is consistent in the sense that it converges in probability to the true log-likelihood function, as the number of particles tends to infinity, see Lemma 4.

Lemma 4. For the model (1)-(3) and $\theta \in \Theta$, the approximate log-likelihood function (61) produced by Algorithm 1 is a consistent estimator of the true log-likelihood (8),

$$
\begin{equation*}
\tilde{\ell}_{T}^{N}(\theta) \xrightarrow{P} \quad \ell_{T}(\theta), \tag{62}
\end{equation*}
$$

as $N \rightarrow \infty$.
In addition to producing an approximation of the intractable log-likelihood function (8), we apply the approximation (60) of the intractable smoothing problem in (27) to produce approximations of the sample score and observed Information matrix via Fisher's and Louis' identities in Lemma 2.

### 6.2. The Approximate Sample Score and Observed Information Matrix

We showed in Section 5 that the sample score and observed Information matrix can be expressed in terms of smoothing problems of the form (27). Appealing to Fisher's identity (37) in Lemma 2, and to the approximation of the smoothing problem (60), we define the particle filter-based approximate sample score as,

$$
\begin{equation*}
\tilde{S}_{T}^{N}(\theta):=\sum_{i=1}^{N} U_{T}\left(\tilde{x}_{1: T}^{(i)} ; \theta\right) \bar{w}_{T}^{(i)} \tag{63}
\end{equation*}
$$

for any parameter $\theta \in \Theta$, with the function $U_{T}\left(x_{1: T} ; \theta\right)$ as defined in (33). If Conjecture 2 holds, then the approximate sample score in (63) is both consistent and asymptotically normal.

Lemma 5. If Conjecture 2 holds and $\theta \in \Theta$, then the approximate sample score (63) is asymptotically normal,

$$
\begin{equation*}
\sqrt{N}\left\{\tilde{S}_{T}^{N}(\theta)-S_{T}(\theta)\right\} \quad \xrightarrow{D} \quad N\left(0, \tilde{\mathbb{S}}_{T}\left[U_{T}\left(x_{1: T} ; \theta\right)\right]\right) \tag{64}
\end{equation*}
$$

as $N \rightarrow \infty$. An intractable expression for the asymptotic covariance matrix $\tilde{\mathbb{S}}_{T}\left[U_{T}\left(x_{1: T} ; \theta\right)\right]$ is given in Lemma A. 5 by setting $t=T$ and $\gamma_{T}\left(x_{1: T}\right)=U_{T}\left(x_{1: T} ; \theta\right)$.

Similarly, by appealing to Louis' identity (38) in Lemma 2, and to the approximation of the smoothing problem (60), we define the particle filter-based approximate observed Information matrix as,

$$
\begin{equation*}
\tilde{I}_{T}^{N}(\theta):=\tilde{S}_{T}^{N}(\theta) \tilde{S}_{T}^{N}(\theta)^{\prime}-\tilde{G}_{T}^{N}(\theta)-\tilde{K}_{T}^{N}(\theta) \tag{65}
\end{equation*}
$$

for any parameter $\theta \in \Theta$, where we define the approximations to (39) and (40) as

$$
\begin{align*}
& \tilde{G}_{T}^{N}(\theta):=\sum_{i=1}^{N} V_{T}\left(\tilde{x}_{1: T}^{(i)} ; \theta\right) \bar{w}_{T}^{(i)}  \tag{66}\\
& \tilde{K}_{T}^{N}(\theta):=\sum_{i=1}^{N} U_{T}\left(\tilde{x}_{1: T}^{(i)} ; \theta\right) U_{T}\left(\tilde{x}_{1: T}^{(i)} ; \theta\right)^{\prime} \bar{w}_{T}^{(i)}, \tag{67}
\end{align*}
$$

and the functions $U_{T}\left(x_{1: T} ; \theta\right)$ and $V_{T}\left(x_{1: T} ; \theta\right)$ are defined in (33) and (34), respectively. If Conjecture 2 holds, then the approximate observed Information in (65) is consistent, stated in the following lemma.

Lemma 6. If Conjecture 2 holds and $\theta \in \Theta$, then the approximate observed Information matrix (65) is consistent,

$$
\begin{equation*}
\tilde{I}_{T}^{N}(\theta) \quad \xrightarrow{P} \quad \boldsymbol{I}_{T}(\theta) \tag{68}
\end{equation*}
$$

as $N \rightarrow \infty$.

Both the approximate sample score (63) and observed Information matrix (65) are biased for finite $N$. This is a general issue related to the particle filter-based approximation of the smoothing problem (60). At each iteration, the particle filter in Algorithm 1 relies on an approximation of the normalized constant, i.e., likelihood contribution. This induces a finite-sample bias in (60) that gradually disappears as the number of particles $N$ tends to infinity and is negligible for large enough $N$, cf. e.g., Robert and Casella (2010, sct. 3.3.2).

The particle filter-based approximation of the sample score (63) and observed Information matrix (65) correspond to a batch version of Algorithm A in Poyiadjis et al. (2011), which is of computational cost $O(N)$, but exhibits quadratically increasing variance of the approximate sample score as a function of the sample size $T$. We note that Poyiadjis et al. (2011) also suggest an alternative algorithm, that exhibits linearly increasing variance as a function of $T$, but at the computational cost $O\left(N^{2}\right)$. For smaller sample sizes, such as monthly observations as usually encountered in economics, we have found that the $O(N)$ algorithm is adequate.

## 7. Particle Filter-Based Estimation and Inference

In this section, we show how the approximate sample score (63) and observed Information matrix (65) can be used to perform parameter estimation and inference. We apply a stochastic approximation method based on the approximate sample score to approximate the ML estimator (19). This has recently been suggested in Poyiadjis et al. (2011). We then use the approximate observed Information matrix to obtain approximate standard errors for the approximate ML estimates. Although these quantities are 'approximate', we note that they can be made arbitrarily precise by increasing the number of particles, $N$, at the expense of increased computational effort.

Recall from Section 4 that the ML estimator (19) is doubly intractable. Consequently, we cannot apply gradient-based optimization algorithms to maximize the log-likelihood function (8). Originally proposed in Robbins and Monro (1951), stochastic approximation methods are conceptually similar to gradient-based optimization methods, but rely on noisy rather than exact evaluations of the sample score to optimize the objective function. The basic idea is that appropriately decreasing the step sizes provides an averaging of the random errors induced by the noisy evaluations of the sample score. For a book-length treatment of stochastic approximation, we refer to Kushner and Yin (2003).

The stochastic approximation algorithm proposed in Poyiadjis et al. (2011, sct. 3.1) consists of a recursion that is conceptually similar to the steepest descent method, cf. e.g., Nocedal and Wright (2006, chp. 3). Prior to executing the algorithm, we choose a fixed initial parameter value $\theta_{0} \in \Theta$, a sequence of particle counts $\left\{N_{j}\right\}_{j=1}^{\infty}$, a sequence of step sizes $\left\{\gamma_{j}\right\}_{j=1}^{\infty}$, and a sequence of weight matrices $\left\{B_{j}\right\}_{j=1}^{\infty}$. The particle counts must be monotonically increasing positive integers, the step sizes must be strictly positive, non-summable but square summable,

$$
\begin{equation*}
\sum_{j=1}^{\infty} \gamma_{j}=\infty \quad \text { and } \quad \sum_{j=1}^{\infty} \gamma_{i}^{2}<\infty \tag{69}
\end{equation*}
$$

and the weight matrices must be positive definite. Having chosen the initial parameter, particle counts, step sizes, and weight matrices, we run the recursion,

$$
\begin{equation*}
\theta_{j+1}=\theta_{j}+\gamma_{j} B_{j} \tilde{S}_{T}^{N_{j}}\left(\theta_{j}\right) \tag{70}
\end{equation*}
$$

for $j=0,1, \ldots, K$. Here $K$ has to be sufficiently large in the sense that the sequence of parameter values generated by the recursion (70) has stabilized in a neighborhood of the true ML estimate. Additionally, if the particle count $N_{j}$ is large enough, the approximation error affecting the stochastic approximation recursion (70) will be approximately normal, cf. Lemma 5 . In this case large disturbances will be rare, such that the parameter sequence $\left\{\theta_{j}\right\}_{j=1}^{K}$ is likely to stabilize without exhibiting large jumps.

We denote by $\left\{\tilde{x}_{1: T, j}^{(i)}, \bar{w}_{t, j}^{(i)}\right\}_{i=1}^{N_{j}}$ the particle paths produced by the particle filter in Algorithm 1 at iteration $j$ of the stochastic approximation recursion (70). The iteration index $j$ is notationally identical to time index of the particle path, cf. Remark 7. Although this is abuse of notation, it is clear from the context whether we refer to the parameter iteration or particle path time index. The parameter $\theta_{j+1}$ produced by iteration $j$ of (70) is a random variable that is measurable with respect to the sub- $\sigma$-algebra $\mathcal{G}_{j}$, defined next.

Definition 2. Let $\mathcal{F}_{T, j}:=\sigma\left(x_{1: T, j}^{(1)}, \ldots, x_{1: T, j}^{\left(N_{j}\right)}\right)$ denote the sub- $\sigma$-algebra in Definition 1 generated with the parameter value $\theta_{j}$, and define the sub- $\sigma$-algebras $\mathcal{G}_{j}:=\mathcal{G}_{j-1} \cup \mathcal{F}_{T, j}$ for $j=1, \ldots$, initialized by $\mathcal{G}_{0}:=\mathcal{F}_{T, 0}$.

One of the main benefits of the stochastic approximation method is that the method is known to stabilize for a wide variety of initial values, sample counts, step sizes, and weight matrices. In practice, all of these choices affect the number of iterations needed to bring the parameter sequence into the neighborhood of the true ML estimator. The choice of step sizes is particularly important, since large step sizes generally speed up the convergence, but fail to dampen the approximation-induced noise. Small step sizes reduce the noise, but cause slow convergence. The particle count has a similar effect, since a low number of particles will result in a computationally cheap but noisy approximation of the sample score, while a large number of particles reduces the noise but increases the computational cost. Heuristically, it is appropriate to use a combination of large step sizes and small particle counts until the parameter sequence has reached a neighborhood of the ML estimator, and then switch to a combination of smaller step sizes and larger particle counts to reduce the noise. The intuition is that, while far away from the ML estimator, a relatively noisy approximation of the sample score will still on average lead the algorithm in the right direction.

The presence of noise in the sample score is not an impediment when applying stochastic approximation, since the use of decreasing step sizes provides an averaging of the errors. However, the finite sample bias of the particle filter-based approximate sample score, cf. Section 6.2, poses a problem since its effect is not mitigated by decreasing the step sizes. Bias reduction is possible by increasing the particle count $N_{j}$ together with the iteration number $j$.

The stochastic approximation method is presented in Algorithm 2 below. ${ }^{1}$

```
Algorithm 2: Stochastic Approximation.
    Choose the initial parameter \(\theta_{0} \in \Theta\), the particle counts \(\left\{N_{j}\right\}_{j=1}^{\infty}\), the step sizes \(\left\{\gamma_{j}\right\}_{j=1}^{\infty}\) and
    weighting matrices \(\left\{B_{j}\right\}_{j=1}^{\infty}\). For \(j=0,1, \ldots, K\) :
```

    1. Run Algorithm 1 for \(\theta_{j}\) to generate \(N_{j}\) weighted particle paths, denoted \(\left\{x_{1: T, j^{\prime}}^{(i)} \bar{w}_{t, j}^{(i)}\right\}_{i=1}^{N_{j}}\).
    2. Compute the approximate sample score (63), denoted

$$
\begin{equation*}
\tilde{S}_{T}^{N_{j}}\left(\theta_{j}\right)=\sum_{i=1}^{N_{j}} U_{T}\left(x_{1: T, j}^{(i)} ; \theta_{j}\right) \bar{w}_{t, j}^{(i)} . \tag{71}
\end{equation*}
$$

3. With step size $\gamma_{j}$, ascend along the direction $B_{j}$,

$$
\begin{equation*}
\theta_{j+1}=\theta_{j}+\gamma_{j} B_{j} \tilde{S}_{T}^{N_{j}}\left(\theta_{j}\right) \tag{72}
\end{equation*}
$$

[^30]Polyak (1990) and Polyak and Juditsky (1992) showed that if the step sizes $\left\{\gamma_{j}\right\}_{j=1}^{\infty}$ satisfy the summability conditions (69) and tend to zero slower than $j^{-1}$, then the average of the last $j-K_{0}$ iterations converges at an optimal rate. Here $K_{0}<K$ denotes the iteration number at which the averaging begins; implicitly, we discard the initial $K_{0}$ iterations. We define the approximate ML estimator as,

$$
\begin{equation*}
\tilde{\theta}_{T}:=\frac{1}{K-K_{0}} \sum_{j=K_{0}}^{K} \theta_{j} \tag{73}
\end{equation*}
$$

suppressing the dependence on the particle count. Establishing convergence of the approximate ML estimator (73) to the true ML estimator (19) is outside the scope of this paper. However, if (73) converges in probability to (19) for any fixed $T$, then (73) inherits the consistency property, cf. Theorem 2 , of the true ML estimator.

Convergence of the particle filter-based stochastic approximation method proposed in Poyiadjis et al. (2011) has, to the author's knowledge, not been studied yet. The finite-sample bias of the approximate sample score (63) presents the primary obstacle to establishing convergence results. Intuition suggests that increasing the number of particles $N_{j}$ with the iteration number $j$ solves the problem. However, convergence of such schemes has not been carefully established, cf. Douc et al. (2014, sct. 12.1.2). Poyiadjis et al. (2011) report stabilization of the particle filter-based stochastic approximation method with constant particle count. In Section 10, we report similar stabilization with increasing particle counts.

If the model is correctly specified, we would conduct inference on the ML estimator via the observed Information matrix, cf. Section 4. Analogously, since the approximate observed Information matrix (65) converges in probability to the true observed Information matrix (10), we can conduct inference for the approximate ML estimator (73) via the approximate observed Information matrix (65), the same way we would conduct inference given the true observed Information matrix (10).

## 8. Model Diagnostics

In this section, we introduce a method to conduct model diagnostics, such that we may assess whether the SSR model is well-specified for a given parameter $\theta$ and observation sequence $y_{1: T}$. Recall that the disturbances $u_{t}, \eta_{t}$ and $v_{t}$ are normally distributed and serially independent with mean zero and unit variances. Because the components $\varepsilon_{t}$ and $\xi_{t}$ are hidden to us, we cannot directly compute the residuals corresponding to the disturbances. Instead, we introduce the normalized one-step prediction errors, cf. Durbin and Koopman (2012, sct. 2.12), that can be approximated via particle filtering. This approach to model diagnostics for state space models has also previously been considered in Pitt and Shephard (1999a).

We define the normalized one-step prediction errors as,

$$
\begin{equation*}
\boldsymbol{e}_{t}:=\operatorname{Var}_{\theta}\left[y_{t} \mid y_{1: t-1}\right]^{-1 / 2}\left(y_{t}-\mathbb{E}_{\theta}\left[y_{t} \mid y_{1: t-1}\right]\right), \tag{74}
\end{equation*}
$$

for $t=1, \ldots, T$. For a well-specified model, the sequence of normalized one-step prediction errors should be serially independent with mean zero with unit variance. Any deviation from these characteristics are indicative of model misspecification.

The conditional mean and variance in (74) can be stated in terms of smoothing problems, where the test functions are the conditional mean and variance of the predictive observation density,

$$
\begin{gather*}
\mathbb{E}_{\theta}\left[y_{t} \mid y_{1: t-1}\right]=\mathbb{E}_{\theta}\left[\mathbb{E}_{\theta}\left[y_{t} \mid x_{t-1}\right] \mid y_{1: t-1}\right]  \tag{75}\\
\operatorname{Var}_{\theta}\left[y_{t} \mid y_{1: t-1}\right]=\mathbb{E}_{\theta}\left[\operatorname{Var}_{\theta}\left[y_{t} \mid x_{t-1}\right] \mid y_{1: t-1}\right]+\operatorname{Var}_{\theta}\left[\mathbb{E}_{\theta}\left[y_{t} \mid x_{t-1}\right] \mid y_{1: t-1}\right] \tag{76}
\end{gather*}
$$

We note that the conditional mean and variance of the predictive observation density are given in Lemma 3. Using the locally optimal particle filter in Algorithm 1, we define approximations to (75) and (76) as

$$
\begin{gather*}
\tilde{\mathbb{E}}_{\theta}^{N}\left[y_{t} \mid y_{1: t-1}\right]:=\sum_{i=1}^{N} \tilde{\mu}_{t \mid t-1}^{y^{\prime}(i)} \bar{w}_{t-1}^{(i)}  \tag{77}\\
\widetilde{\operatorname{Var}}_{\theta}^{N}\left[y_{t} \mid y_{1: t-1}\right]:=\sum_{i=1}^{N} \tilde{\Sigma}_{t \mid t-1}^{y_{,}(i)} \bar{w}_{t-1}^{(i)}+\sum_{i=1}^{N}\left(\tilde{\mu}_{t \mid t-1}^{y,(i)}\right)\left(\tilde{\mu}_{t \mid t-1}^{y^{\prime}(i)}\right)^{\prime} \bar{w}_{t-1}^{(i)}-\tilde{\mathbb{E}}_{\theta}^{N}\left[y_{t} \mid y_{1: t-1}\right] \tilde{\mathbb{E}}_{\theta}^{N}\left[y_{t} \mid y_{1: t-1}\right]^{\prime}, \tag{78}
\end{gather*}
$$

respectively, where we have defined the conditional moments given each individual particle as,

$$
\begin{align*}
& \tilde{\mu}_{t \mid t-1}^{y_{( }^{\prime}(i)}:=\mathbb{E}_{\theta}\left[y_{t} \mid \tilde{x}_{t-1}^{(i)}\right]  \tag{79}\\
& \tilde{\Sigma}_{t \mid t-1}^{y_{( }(i)}:=\operatorname{Var}_{\theta}\left[y_{t} \mid \tilde{x}_{t-1}^{(i)}\right] \tag{80}
\end{align*}
$$

for $i=1, \ldots, N$. Finally, we use the approximations (77) and (78) to define the approximate normalized likelihood contributions as follows,

$$
\begin{equation*}
\tilde{e}_{t}^{N}:=\widetilde{\operatorname{Var}}_{\theta}^{N}\left[y_{t} \mid y_{1: t-1}\right]^{-1 / 2}\left(y_{t}-\widetilde{\mathbb{E}}_{\theta}^{N}\left[y_{t} \mid y_{1: t-1}\right]\right) \tag{81}
\end{equation*}
$$

for $t=1, \ldots, T$. Thus, by applying the particle filter in Algorithm 1, we obtain the sequence of approximate normalized one-step prediction errors $\tilde{e}_{1: T}$ via (77)-(81). For $N$ sufficiently large, we can use the sequence $\tilde{e}_{1: T}$ to test whether the true sequence of normalized one-step prediction errors $\boldsymbol{e}_{1: T}$ is serially independent with mean zero and unit variance. For common tests for serial dependence and ARCH effects see e.g., Doornik and Hendry (2013, sct. 11.9.2-3).

## 9. Simulation Study

In this section, we conduct a simulation study of the asymptotic properties of the ML estimator, stated in Theorem 2. We limit our treatment to $B, A, \Phi$ and $\Omega_{\Phi}$, leaving aside the remaining parameters $\Omega_{u}, \mu, \Omega_{\eta}, \Omega_{v}$ and $\Omega_{\eta, v}$. Recall, the loading matrix for the stationary components $A$ is conjectured to be asymptotically normal, while the loading matrix of the nonstationary components $B$ is kept fixed. Due to the results of Chang et al. (2009), we expect the asymptotic distribution of $B$ to be mixed normal, and we tentatively investigate this. Moreover, we consider the case where $\Phi_{t}$ is a stochastic unit root. A deterministic unit root is associated with the Dickey-Fuller distribution, cf. Dickey and Fuller (1979), while a stochastic unit root has been shown to be asymptotically normal, see e.g., Ling (2007) and Bohn Nielsen and Rahbek (2014).

Recall, Theorem 2 is based on the conjectured properties of the true, intractable log-likelihood function and its derivatives, cf. Conjecture 1. The aim is to substantiate this conjecture by obtaining the distribution of the approximate ML estimator based on simulated data sets. Usually, the number of realizations in a simulation study of this type is in excess of 1000 and the sample length in excess of 2500 observations. Due to the computational intensity of the particle filter-based stochastic approximation method in Algorithm 2, we limit ourselves to 250 realizations and 500 observations.

We let each of the simulated data sets be a bivariate $p=2$ series of length $T=500$ observations with $r=1$ stationary component and $p-r=1$ nonstationary component. We use the parameter

$$
\begin{gather*}
B=\left[\begin{array}{l}
1 \\
1
\end{array}\right], \quad A=\left[\begin{array}{l}
0 \\
1
\end{array}\right], \quad \Omega_{u}=\left[\begin{array}{cc}
2.5^{2} & 0 \\
0 & 2.5^{2}
\end{array}\right],  \tag{82}\\
\mu=0, \quad \phi=1, \quad \omega_{\phi}^{2}=0.25^{2}, \quad \omega_{\eta}^{2}=15^{2}, \quad \omega_{\eta, v}=0, \quad \text { and } \quad \omega_{v}^{2}=2.5^{2}, \tag{83}
\end{gather*}
$$

to generate the simulated data sets. We note the parameter values (83) result in a top Lyapunov coefficient of $\gamma_{n}=-0.035$, computed via (15) with $n=10^{6}$, such that the RCAR process $\left\{\xi_{t}\right\}_{t=0,1, \ldots}$ is strictly stationary.

Having simulated 250 series with the data generating process given by (1)-(3) and (83), we apply Algorithm 2 with $K=600$ iterations to obtain the approximate ML estimate for the parameter in question, e.g., $\phi$, keeping all other parameters fixed at the true values in (83). We initialize the algorithm at the true parameter value, and initiate Polyak averaging at iteration $K_{0}=100 .{ }^{2}$ Moreover, we let the particle count increase as

$$
\begin{equation*}
N_{j}=50+\lfloor 1 / 20 j\rfloor, \tag{84}
\end{equation*}
$$

where $\lfloor\cdot\rfloor$ denotes the largest integer that is smaller than the argument. We let the step size sequence to decrease as

$$
\begin{equation*}
\gamma_{j}=100(j+500)^{-2 / 3} \tag{85}
\end{equation*}
$$

and set the weight matrix to

$$
B_{j}=T^{-1} \operatorname{diag}\left(\left[\begin{array}{llllllllll}
10^{-5} & 1 & 1 & 1 & 1 & 10^{-2} & 1 & 1 & 1 & 10^{-3} \tag{86}
\end{array}\right]\right)
$$

for $j=1,2, \ldots, K$. Note the particle count (84) tends to infinity as $j \rightarrow \infty$, eliminating the finite-sample bias of (63)-(65), the step sizes satisfy (69), and the weight matrix is constant. ${ }^{3}$

The results from the simulation experiment are presented in Figure 1. Despite the relatively low number of realizations and observations, Figure 1 is instructive of the asymptotic distributions of $A_{1}, \phi$ and $\omega_{\phi}^{2}$, cf. Panels (a), (c) and (d). These all appear to be normal. Recall, Theorem 2 does not state the asymptotic distribution of the ML estimator for $B_{2}$, and from Panel (b) it does not appear to be normal. Rather, the realizations in Panel (b) are consistent with mixed normality, as we would expect from the closely-related CST model, cf. Chang et al. (2009). To investigate further, one could to simulate the $t$-ratios of $B_{2}$, which should be standard normal. This involves the approximation of the observed Information matrix for each realization, which further increases the computational cost. For this reason, and because we consider $B$ fixed, we do not pursue this further here.

[^31]

Figure 1. Simulation study with 250 realizations of the approximate MLE for $A_{1}, B_{2}, \phi$, and $\omega_{\phi}^{2}$.

In summary, the findings of the simulation study tentatively support the conjecture made in Section 4. Namely, the ML estimator for $A, \Phi$ and $\Omega_{\Phi}$ is asymptotically normal. The ML estimator for $B_{2}$ appears to be consistent with mixed normality. We have not investigated the remaining parameters.

## 10. An Illustration

In this section, we illustrate the use of the SSR model by applying it to the monthly 10-year government bond rates for Germany and Greece from January 1999 to February 2018. ${ }^{4}$ We denote the German and Greek bond rates $y^{G E}$ and $y^{G R}$, respectively, and measure these in basis points per year. The sample begins at the introduction of the euro area and ends at present day. During this period, the rates initially exhibit convergence towards a common 'euro area rate', until interrupted by the euro area crisis beginning in 2009 and culminating in 2011. The rates, the spread and the changes in the spread are illustrated in Figure 2 below. Because the spread is up to 75 times larger during the second half of the sample than during the first half, we split the display of the sample into the first and second half, respectively.

Panels (a) and (b) in Figure 2 show the bond rates, Panels (c) and (d) show the spread, and Panels (e) and (f) show the changes in the spread in the two periods. We note two features of the observations. First, Panel (a) suggests the rates can be characterized by a shared common stochastic trend, since these tend to move in tandem. Second, Panels (d) and (f) suggest the spread can be characterized by a RCAR process, since the changes in the spread, cf. Panel (f), are clearly positively associated with the level of the spread itself, cf. Panel (d).

[^32]

Figure 2. German and Greek 10-year government bond rates, spread and changes in the spread. Monthly observations in basis points from January 1999 to February 2018.

We define the observation vector as $y_{t}:=\left[\begin{array}{ll}y_{t}^{G E} & y_{t}^{G R}\end{array}\right]^{\prime}$. We condition on the observation for January 1999, which we denote $y_{0}$, such that the effective sample spans $t=1, \ldots, 229$. From visual inspection of Figure 2, our working assumption is that the spread $y_{t}^{G R}-y_{t}^{G E}$ is strictly stationary, while the rates $y_{t}$ share a common stochastic trend. With a $p=2$ dimensional system, we thus have $r=1$ stationary component and $p-r=1$ nonstationary component. Moreover, we fix $B=\left[\begin{array}{ll}1 & 1\end{array}\right]^{\prime}$, such that the orthogonal complement $b=\left[\begin{array}{cc}-1 & 1\end{array}\right]^{\prime}$ produces the spread. To ensure the model is just-identified, we normalize on the second element of $A$, such that $A_{2}=1$.

We apply the particle filter-based stochastic approximation method in Algorithm 2 to obtain the approximate ML estimate of the model parameter $\theta$. For this illustration, we run the algorithm for $K=10,000$ iterations. We let the particle count increase as (84), the step size sequence decrease as (85), and the weighting matrix as (86). We initiate Polyak averaging at iteration $K_{0}=5000$.


Figure 3. Parameter and log-likelihood sequences from stochastic approximation with $K=10,000$ iterations. We also show a moving average of lag order 500 for the log-likelihood sequence. To avoid large differences in the scales of the displayed sequences, we have scaled the sequences for $A_{1}, \omega_{\eta}^{2}$, $\omega_{\eta, v}, \omega_{v}$, and $\omega_{\phi}$ by $100,1 / 300,1 / 50,1 / 50$, and 2 respectively.

Figure 3 shows the results of running the particle filter-based stochastic approximation method. Panel (a) displays the iterations for the parameters in the observation Equation (22), Panel (b) displays the iterations for the parameters in the transition Equation (23), and Panel (c) displays the sequence of realized approximate log-likelihoods together with a moving average of lag order 500. The algorithm has been implemented in the Ox 7 programming language, cf. Doornik (2012), using analytical derivatives of the complete data log-likelihood (32) for the evaluation of the function (33). The elements of the parameter sequence shown in Panels (a) and (b) have stabilized after the initial 7500 iterations. At the 10,000th iteration, the particle count has increased to 550 , the step size decreased to 0.2085 , and the sequences have stabilized. By inspection of the sequence of the approximate log-likelihood in Panel (c), we see that the value has also stabilized after approximately 7500 iterations.

The estimation results are presented in Table 1, together with approximate classic standard errors. ${ }^{5}$ Before considering inference, we assess the model fit. We compute the normalized one-step prediction errors $\tilde{e}_{1: T}^{N}$ via (81) using $N=1000$ particles. Table 2 presents univariate tests for autocorrelation (AR) of order one and two, autoregressive conditional heteroskedasticity (ARCH) of order one, and a multivariate test for AR of order one and two, cf. Doornik and Hendry (2013, sct. 11.9.2-3). We cannot reject the null hypothesis of no-AR of order one and two in the univariate as well as multivariate tests at a $5 \%$ critical level. Nor can we reject the null hypothesis of no-ARCH for the residuals at a $5 \%$ critical level. However, we note the test for the German rate is close to, but below, our chosen critical level. This could suggest unmodeled heteroskedasticity in the German bond rate. In conclusion, the overall specification of the model is acceptable. Moreover, computing the top Lyapunov coefficient via (15) with $n=10^{5}$ produces a coefficient of $\hat{\gamma}_{n}=-0.007$, which indicates the stationary direction is strictly stationary for $\tilde{\theta}_{T}$.

Table 1. Approximate ML estimate, $\tilde{\theta}_{T}$.

| Parameter | Estimate | Std.err. | Parameter | Estimate | Std.err. |
| :---: | ---: | :---: | :---: | ---: | :---: |
| $B_{1}$ | 1.0000 | - | $\mu$ | 0.3449 | 0.5526 |
| $B_{2}$ | 1.0000 | - | $\phi$ | 1.0085 | 0.0152 |
| $A_{1}$ | -0.0154 | $3.4 \times 10^{-5}$ | $\omega_{\phi}^{2}$ | 0.0306 | 0.0031 |
| $A_{2}$ | 1.0000 | - | $\omega_{\eta}^{2}$ | 360.1600 | 33.6250 |
| $\omega_{u 11}^{2}$ | 2.1063 | 0.1974 | $\omega_{\eta, v}$ | -22.2400 | 0.8119 |
| $\omega_{u 12}$ | 3.5924 | 0.3435 | $\omega_{v}^{2}$ | 1.7880 | 2.0728 |
| $\omega_{u 22}^{2}$ | 6.6327 | 0.6214 |  |  |  |

Note: The approximate log-likelihood is $\tilde{\ell}_{T}=-2094.1$. The approximate ML estimate has been obtained by running Algorithm 2 for $K=10,000$ iterations with the particle count increasing to $N=550$ particles, as described in the main text. The standard errors are based on the inverse of the approximate observed Information matrix computed with $N=1000$ particles.

Table 2. Model diagnostics.

| Univariate tests for AR 1-2: | $\tilde{e}_{t, 1}$ | $F(2,227)=1.4523$ | $p=0.2362$ |
| :--- | :--- | :--- | :--- |
|  | $\tilde{e}_{t, 2}$ | $F(2,227)=1.2086$ | $p=0.3005$ |
| Multivariate test for AR 1-2: |  | $F(8,448)=1.6084$ | $p=0.1200$ |
|  | $\tilde{e}_{t, 1}$ | $F(1,227)=4.7008$ | $p=0.0312$ |
|  | $\tilde{e}_{t, 2}$ | $F(1,227)=0.58861$ | $p=0.4438$ |

Note: The approximate normalized one-step prediction errors $\tilde{e}_{1: T}$ have been computed with $N=$ 1000 particles for the approximate ML estimate $\tilde{\theta}_{T}$, cf. Table 1.

The model is reasonably well-specified, and we therefore proceed to use the approximate classic standard errors to conduct inference on the approximate ML estimates. First, we note the standard error of the estimate of $A_{1}$ is extremely small. Since the test for no-ARCH for the residuals associated with the German rate is rejected at the $5 \%$ critical level, this could affect the approximate classic standard errors. ${ }^{6}$ Nevertheless, it is economically plausible that the stationary component also loads into the German rate, given that a large increase in the Greek rate would in this case coincide with a small drop in the German rate, which is consistent with risk-averse investors seeking safer assets in times of uncertainty, such as the euro area crisis. Second, we cannot reject the null hypothesis that $H_{0}: \phi=1$ at a $5 \%$ critical level with $p=0.577$. Third, the estimate of $\omega_{\phi}^{2}$ is significantly different from

[^33]zero at any commonly used critical level. However, the constant term $\mu$ is not significantly different from zero with $p=0.533$. Fourth, the measurement errors are highly positively correlated with coefficient 0.961 , and the innovations of the unobserved components are highly negatively correlated with coefficient -0.876 . The results in Table 1 suggest the level of the stationary direction is a stochastic unit root process without a constant term. An approximate likelihood ratio test for the joint null hypothesis $H_{0}: \phi=1, \mu=0$ fails to reject the null at a $5 \%$ critical level with $p=0.374$.

Based on the estimates in Table 1, we use the orthogonal complements $b$ and $a$ to compute the changes of the nonstationary and stationary components, given by $b^{\prime} \Delta y_{t}$ and $a^{\prime} \Delta y_{t}$, respectively. These are illustrated in Figure 4. First, we note the magnitude of the changes in Panels (a) and (b) of Figure 4 are slightly larger during the second half of the sample than during the first (standard deviations 18.01 and 20.16, respectively). Otherwise, the series in Panels (a) and (b) in Figure 4 are consistent with a homoskedastic random walk plus measurement error, cf. (17). The magnitude of the changes in Panels (c) and (d) of Figure 4 is positively associated with the level, just as observed in Figure 2. This is consistent with a random coefficient autoregressive process plus measurement error, cf. (18).


Figure 4 . Changes in the nonstationary $b^{\prime} y_{t}$ and stationary $a^{\prime} y_{t}$ components.

Summarizing, the empirical illustration suggests that the SSR model successfully characterizes the 10-year government bond rates for Germany and Greece during the period from January 1999 to February 2018. During this sample, the spread exhibits bubble-like behavior, which is captured by the random coefficient autoregressive dynamics of the stationary component. Additionally, the levels exhibit a shared common stochastic trend, which is captured by the random walk dynamics of the nonstationary component.

## 11. Conclusions

In this paper, we have proposed and studied the stochastic stationary root model, which is a multivariate nonlinear state space model. We introduced particle filter-based approximations of the intractable log-likelihood function, sample score and observed Information matrix. In turn, we used
these to approximate the ML estimator via stochastic approximation, and showed how to perform inference via the approximate observed Information matrix. We considered model diagnostics to assess the model fit. Additionally, we conducted a simulation study to investigate the asymptotic properties of the ML estimator. Finally, we presented an empirical application to the 10-year government bond rates in Germany and Greece in the period from January 1999 to February 2018 to illustrate the usefulness of the SSR model.

Acknowledgments: The author gratefully acknowledges comments by two anonymous referees that have led to substantial improvements of the paper. The author also thanks the editors, Rocco Mosconi and Paolo Paruolo, for constructive feedback, and the assistant editor, Lu Liao, for assisting in the publication process. Finally, the author would like to thank Anders Rahbek, Michael Pitt, Siem Jan Koopman, Heino Bohn Nielsen, Katarina Juselius, Søren Johansen, Simon Hetland, Gareth Roberts, Adam Johansen, Axel Finke, and Anthony Lee for helpful comments and dicsussions. Part of the work was undertaken while the author was a PhD student at the Department of Economics at the University of Copenhagen and part of the work was undertaken while the author was a CRiSM Research Fellow at the Department of Statistics at the University of Warwick. While at the University of Warwick, funding from the 36 Engineering and Physical Sciences Research Council (EPSRC) is gratefully acknowledged (Grant EP /D002060/1). All errors and omissions are the sole responsibility of the author.

Conflicts of Interest: The author declares no conflict of interest.

## Appendix A. Auxiliary Results

Lemma A.1. For the SSR model (1)-(3) with $\theta \in \Theta$, it holds that
$i \quad \int p_{\omega}\left(y_{t} \mid x_{t}\right) p_{\lambda}\left(x_{t} \mid x_{t-1}\right) \mathrm{d} x_{t}>0$ for all $x_{t-1} \in \mathcal{R}^{p}$, and
ii $\sup _{x_{t} \in \mathcal{R}^{p}} p_{\omega}\left(y_{t} \mid x_{t}\right)<\infty$,
for any $t \in\{1, \ldots, T\}$.
Proof of Lemma A.1. By Corollary 1 we have that $\int p_{\omega}\left(y_{t} \mid x_{t}\right) p_{\lambda}\left(x_{t} \mid x_{t-1}\right) \mathrm{d} x_{t}=p_{\theta}\left(y_{t} \mid x_{t-1}\right)$ is Gaussian, and therefore strictly positive for all $x_{t-1} \in \mathcal{R}^{p}$ and $\theta \in \Theta$, which yields part (i). Moreover, because the observation density (25) is Gaussian with constant and non-singular covariance matrix, we obtain part (ii).

Lemma A.2. For the SSR model (1)-(3) with $\theta \in \Theta$, the model likelihood $\boldsymbol{p}_{\theta}\left(y_{1: T}\right)$ is strictly positive and finite,

$$
\begin{equation*}
0<\boldsymbol{p}_{\theta}\left(y_{1: T}\right)<\infty . \tag{A1}
\end{equation*}
$$

Proof of Lemma A.2. Preliminarily, we observe the likelihood in (A1) can equivalently be written in terms of the complete data likelihood $p_{\theta}\left(y_{1: T}, x_{1: T}\right)$,

$$
\begin{equation*}
\boldsymbol{p}_{\theta}\left(y_{1: T}\right)=\int p_{\theta}\left(y_{1: T}, x_{1: T}\right) \mathrm{d} x_{1: T} \tag{A2}
\end{equation*}
$$

which, by the state space structure of the model, cf. (25)-(26), is equivalently

$$
\begin{equation*}
\boldsymbol{p}_{\theta}\left(y_{1: T}\right)=\int \prod_{t=1}^{T} p_{\omega}\left(y_{t} \mid x_{t}\right) p_{\lambda}\left(x_{t} \mid x_{t-1}\right) \mathrm{d} x_{1: T} \tag{A3}
\end{equation*}
$$

By Lemma A.1.(i) and (A3), we have that the likelihood in (A1) is strictly positive, since

$$
\begin{equation*}
\int \prod_{t=1}^{T} p_{\omega}\left(y_{t} \mid x_{t}\right) p_{\lambda}\left(x_{t} \mid x_{t-1}\right) \mathrm{d} x_{1: T}>0 \tag{A4}
\end{equation*}
$$

Moreover, by Lemma A.1.(ii), the likelihood in (A1) is also finite, since

$$
\begin{align*}
& \int \prod_{t=1}^{T} p_{\omega}\left(y_{t} \mid x_{t}\right) p_{\lambda}\left(x_{t} \mid x_{t-1}\right) \mathrm{d} x_{1: T} \\
& \quad \leq \prod_{t=1}^{T} \sup _{x_{t} \in \mathcal{R}^{p}} p_{\omega}\left(y_{t} \mid x_{t}\right) \int \prod_{t=1}^{T} p_{\lambda}\left(x_{t} \mid x_{t-1}\right) \mathrm{d} x_{1: T}  \tag{A5}\\
& \quad=\prod_{t=1}^{T} \sup _{x_{t} \in \mathcal{R}^{p}} p_{\omega}\left(y_{t} \mid x_{t}\right)<\infty,
\end{align*}
$$

which completes the proof of Lemma A. 2.
Lemma A.3. For the model (1)-(3) with $\theta \in \Theta$, it holds that
i $\quad p_{\omega}\left(y_{t} \mid x_{t}\right) p_{\lambda}\left(x_{t} \mid x_{t-1}\right) \ll p_{\theta}\left(x_{t} \mid x_{t-1}, y_{t}\right)$ for all $x_{t-1} \in \mathcal{R}^{p}$,
ii $\sup _{x_{t-1}, x_{t} \in \mathcal{R}^{p} \times \mathcal{R}^{p}} \frac{p_{\omega}\left(y_{t} \mid x_{t}\right) p_{\lambda}\left(x_{t} \mid x_{t-1}\right)}{p_{\theta}\left(x_{t} \mid x_{t-1}, y_{t}\right)}>0$, and
iii $p_{\theta}\left(x_{t} \mid x_{t-1}, y_{t}\right)>0$ for all $x_{t-1} \in \mathcal{R}^{p}$,
for $t \in\{1, \ldots, T\}$
Proof of Lemma A.3. We preliminarily note that the locally optimal transition density (46) can be written as

$$
\begin{equation*}
p_{\theta}\left(x_{t} \mid x_{t-1}, y_{t}\right)=\frac{p_{\omega}\left(y_{t} \mid x_{t}\right) p_{\lambda}\left(x_{t} \mid x_{t-1}\right)}{p_{\theta}\left(y_{t} \mid x_{t-1}\right)} \tag{A6}
\end{equation*}
$$

where the predictive observation density is given by the integral,

$$
\begin{equation*}
p_{\theta}\left(y_{t} \mid x_{t-1}\right)=\int p_{\omega}\left(y_{t} \mid x_{t}\right) p_{\lambda}\left(x_{t} \mid x_{t-1}\right) \mathrm{d} x_{t} \tag{A7}
\end{equation*}
$$

By (A6) and the definition of absolute continuity, part (i) states that for every Borel-measurable set $\mathcal{A} \in \mathcal{B}\left(\mathcal{R}^{p}\right)$, it holds that

$$
\begin{equation*}
\int_{\mathcal{A}} \frac{p_{\omega}\left(y_{t} \mid x_{t}\right) p_{\lambda}\left(x_{t} \mid x_{t-1}\right)}{p_{\theta}\left(y_{t} \mid x_{t-1}\right)} \mathrm{d} x_{t}=0 \quad \Longrightarrow \quad \int_{\mathcal{A}} p_{\omega}\left(y_{t} \mid x_{t}\right) p_{\lambda}\left(x_{t} \mid x_{t-1}\right) \mathrm{d} x_{t}=0 . \tag{A8}
\end{equation*}
$$

By (A7) and Lemma A.1. $(i)$, we know the predictive observation density is strictly positive $p_{\theta}\left(y_{t} \mid x_{t-1}\right)>0$ for all $x_{t-1} \in \mathcal{R}^{p}$ and $\theta \in \Theta$. Therefore (A8) is true for all $x_{t-1} \in \mathcal{R}^{p}$ and $\theta \in \Theta$, and part (i) holds.

To show part (ii), we first use (A6) to write

$$
\begin{equation*}
\frac{p_{\omega}\left(y_{t} \mid x_{t}\right) p_{\lambda}\left(x_{t} \mid x_{t-1}\right)}{p_{\theta}\left(x_{t} \mid x_{t-1}, y_{t}\right)}=p_{\theta}\left(y_{t} \mid x_{t-1}\right), \tag{A9}
\end{equation*}
$$

where, by Corollary 1, we have that $p_{\theta}\left(y_{t} \mid x_{t-1}\right)$ is Gaussian and therefore strictly positive for all $x_{t}, x_{t-1} \in \mathcal{R}^{p} \times \mathcal{R}^{p}$ and $\theta \in \Theta$, and part (ii) holds.

Part (iii) follows from $p_{\theta}\left(x_{t} \mid x_{t-1}, y_{t}\right)$ being Gaussian, cf. Lemma 3, and therefore strictly positive for all $x_{t-1} \in \mathcal{R}^{p}$. Thus, part (iii) holds.

Lemma A.4. If $\theta \in \Theta$ and $\gamma_{t}\left(x_{1: t}\right) \in L^{1}\left[\mathcal{R}^{t p}, \boldsymbol{p}_{\theta}\left(x_{1: t} \mid y_{1: t}\right)\right]$, then it holds that the approximation (60) is consistent,

$$
\begin{equation*}
\tilde{\mathbb{E}}_{\theta}^{N}\left[\gamma_{t}\left(x_{1: t}\right) \mid y_{1: t}\right] \quad \xrightarrow{P} \quad \mathbb{E}_{\theta}\left[\gamma_{t}\left(x_{1: t}\right) \mid y_{1: t}\right] \tag{A10}
\end{equation*}
$$

for any $t \in\{1, \ldots, T\}$, as $N \rightarrow \infty$.
Proof of Lemma A.4. We apply Theorem 9.4.5.(i) in Cappé et al. (2005) by verifying its conditions, i.e., Assumptions 9.4.1-3. We note the theorem is stated for scalar test functions, but generalizes to higher-dimensional test functions. Assumptions 9.4.1-2 is hold by Lemma A.1, while Assumption 9.4.3 holds by Lemma A.3. Thus, the conditions for Theorem 9.4.5.(i) in Cappé et al. (2005) are satisfied, which completes the proof of Lemma A.4.

Lemma A.5. If $\theta \in \Theta$ and $\gamma_{t}\left(x_{1: t}\right) \in L^{2}\left[\mathcal{R}^{t p}, \boldsymbol{p}_{\theta}\left(x_{1: t} \mid y_{1: t}\right)\right]$, then it holds that the approximation (60) is consistent and asymptotically normal,

$$
\begin{equation*}
\sqrt{N}\left\{\tilde{\mathbb{E}}_{\theta}^{N}\left[\gamma_{t}\left(x_{1: t}\right) \mid y_{1: t}\right]-\mathbb{E}_{\theta}\left[\gamma_{t}\left(x_{1: t}\right) \mid y_{1: t}\right]\right\} \quad \xrightarrow{D} \quad N\left(0, \tilde{\mathbb{S}}_{t}\left[\gamma_{t}\left(x_{1: t}\right)\right]\right) \tag{A11}
\end{equation*}
$$

for any $t \in\{1, \ldots, T\}$, as $N \rightarrow \infty$. Initialized by $\tilde{\mathbb{S}}_{0}:=0$, the asymptotic covariance matrix $\tilde{\mathbb{S}}_{t}\left[\gamma_{t}\left(x_{1: t}\right)\right]$ is given by

$$
\begin{align*}
& \tilde{\mathbb{S}}_{t}\left[\gamma_{t}\left(x_{1: t}\right)\right]=\tilde{\mathbb{S}}_{t-1}\left[\mathbb{E}_{q, t}\left[\left.\left(\gamma_{t}\left(x_{1: t}\right)-\mathbb{E}_{\theta}\left[\gamma_{t}\left(x_{1: t}\right) \mid y_{1: t}\right]\right) \frac{\tilde{w}_{t}\left(x_{t-1: t}\right)}{\boldsymbol{p}_{\theta}\left(y_{t} \mid y_{1: t-1}\right)} \right\rvert\, x_{1: t-1}\right]\right] \\
&+\operatorname{Var}_{\theta}\left[\left.\mathbb{E}_{q, t}\left[\left.\left(\gamma_{t}\left(x_{1: t}\right)-\mathbb{E}_{\theta}\left[\gamma_{t}\left(x_{1: t}\right) \mid y_{1: t}\right]\right) \frac{\tilde{w}_{t}\left(x_{1: t}\right)}{\boldsymbol{p}_{\theta}\left(y_{t} \mid y_{1: t-1}\right)} \right\rvert\, x_{1: t-1}\right] \right\rvert\, y_{1: t-1}\right]  \tag{A12}\\
&+\mathbb{E}_{\theta}\left[\left.\operatorname{Var}_{q, t}\left[\left.\left(\gamma_{t}\left(x_{1: t}\right)-\mathbb{E}_{\theta}\left[\gamma_{t}\left(x_{1: t}\right) \mid y_{1: t}\right]\right) \frac{\tilde{w}_{t}\left(x_{1: t}\right)}{\boldsymbol{p}_{\theta}\left(y_{t} \mid y_{1: t-1}\right)} \right\rvert\, x_{1: t-1}\right] \right\rvert\, y_{1: t-1}\right],
\end{align*}
$$

where, for any appropriately integrable function $\gamma\left(x_{1: t}\right)$, we define the operators

$$
\begin{gather*}
\mathbb{E}_{q, t}\left[\gamma\left(x_{1: t}\right) \mid x_{1: t-1}\right]:=\int \gamma\left(x_{1: t}\right) q_{\theta}\left(x_{t} \mid f_{1: x-1}, y_{1: t-1}\right) \mathrm{d} x_{1: t}  \tag{A13}\\
\operatorname{Var}_{q, t}\left[\gamma\left(x_{1: t}\right) \mid x_{1: t-1}\right]:=\mathbb{E}_{q, t}\left[\gamma\left(x_{1: t}\right) \gamma\left(x_{1: t}\right)^{\prime} \mid x_{1: t-1}\right]-\mathbb{E}_{q, t}\left[\gamma\left(x_{1: t}\right) \mid x_{1: t-1}\right] \mathbb{E}_{q, t}\left[\gamma\left(x_{1: t}\right) \mid x_{1: t-1}\right]^{\prime}, \tag{A14}
\end{gather*}
$$

omitting dependence on $\theta$.
Proof of Lemma A.5. We apply Theorem 9.4.5.(ii) in Cappé et al. (2005) by verifying its conditions, i.e., Assumptions 9.4.1-3. Similar to the proof of Lemma A.4, we note the theorem is stated for scalar test functions, but generalizes to higher-dimensional test functions. Assumptions 9.4.1-2 is hold by Lemma A.1, while Assumption 9.4.3 holds by Lemma A.3. Thus, the conditions for Theorem 9.4.5.(ii) in Cappé et al. (2005) are satisfied, which completes the proof of Lemma A.5.

## Appendix B. Main Results

Proof of Lemma 1. We compute conditional mean and variance of $\xi_{t}$ in Equation (2). First the mean

$$
\begin{align*}
\mathbb{E}_{\lambda}\left[\xi_{t} \mid \xi_{t-1}\right] & =\mathbb{E}_{\lambda}\left[\mu+\Phi_{t} \xi_{t-1}+v_{t} \mid \xi_{t-1}\right] \\
& =\mu+\Phi \xi_{t-1} \tag{A15}
\end{align*}
$$

and then the variance

$$
\begin{align*}
\operatorname{Var}_{\lambda}\left[\xi_{t} \mid \xi_{t-1}\right] & =\operatorname{Var}_{\lambda}\left[\mu+\Phi_{t} \xi_{t-1}+v_{t} \mid \xi_{t-1}\right] \\
& =\operatorname{Var}_{\lambda}\left[\mu+\left(\xi_{t-1}^{\prime} \otimes I_{r}\right) \operatorname{vec}\left(\Phi_{t}\right)+v_{t} \mid \xi_{t-1}\right] \\
& =\left(\xi_{t-1}^{\prime} \otimes I_{r}\right) \operatorname{Var}_{\lambda}\left[\operatorname{vec}\left(\Phi_{t}\right)\right]\left(\xi_{t-1}^{\prime} \otimes I_{r}\right)^{\prime}+\operatorname{Var}_{\lambda}\left[v_{t}\right]  \tag{A16}\\
& =\left(\xi_{t-1}^{\prime} \otimes I_{r}\right) \Omega_{\Phi}\left(\xi_{t-1}^{\prime} \otimes I_{r}\right)^{\prime}+\Omega_{v}
\end{align*}
$$

Since the conditional distribution of $\xi_{t}$ given $\xi_{t-1}$ is Gaussian, it is completely characterized by its first and second conditional moments. Thus, we obtain equations (12)-(13), which completes the proof of Lemma 1.

Proof of Lemma 2. The result is an application of the Fisher's and Louis' identities to the SSR model. We use Proposition 10.1.6 in Cappé et al. (2005), by verifying the conditions.

First, we verify that Assumption 10.1.3 in Cappé et al. (2005) holds. We have that $\Theta$ is an open subset of $\mathcal{R}^{d_{\theta}}$, which satisfies Assumption 10.1.3.(i). Assumption 10.1.3.(ii) is satisfied via Lemma A.2. Assumption 10.1.3.(iii) is encompassed by condition (b) of Proposition 10.1.6 in Cappé et al. (2005), shown below. Thus, Assumption 10.1.3 in Cappé et al. (2005) holds.

Second, we verify conditions (a) and (b) of Proposition 10.1.6 in Cappé et al. (2005). Condition (a) holds by Conjecture 1. For condition (b), we begin with the third and last part, which states that

$$
\begin{equation*}
\frac{\partial}{\partial \theta} \int \log p_{\theta}\left(y_{1: t}, x_{1: T}\right) \boldsymbol{p}_{\vartheta}\left(x_{1: T} \mid y_{1: T}\right) \mathrm{d} x_{1: T}=\int \frac{\partial}{\partial \theta} \log p_{\theta}\left(y_{1: t}, x_{1: T}\right) \boldsymbol{p}_{\vartheta}\left(x_{1: T} \mid y_{1: T}\right) \mathrm{d} x_{1: T} \tag{A17}
\end{equation*}
$$

For $\theta, \vartheta \in \Theta$, the complete data log-likelihood (32) is log-Gaussian and therefore continuous with respect to $\theta$, and (A17) holds.

The second part of condition (b) states that for $\theta \in \Theta$,

$$
\begin{align*}
& \int\left\|U_{T}\left(x_{1: T} ; \theta\right)\right\| \boldsymbol{p}_{\theta}\left(x_{1: T} \mid y_{1: T}\right) \mathrm{d} x_{1: T}<\infty  \tag{A18}\\
& \int\left\|V_{T}\left(x_{1: T} ; \theta\right)\right\| \boldsymbol{p}_{\theta}\left(x_{1: T} \mid y_{1: T}\right) \mathrm{d} x_{1: T}<\infty \tag{A19}
\end{align*}
$$

which is holds by Conjecture 2.
The first part of condition (b) states that for $\theta, \vartheta \in \Theta$, the entropy function in (31) is twice-differentiable with respect to $\theta$ for fixed $\vartheta$ and $y_{1: T}$. Using (A17) and that the complete data $\log$-likelihood (32) is twice-differentiable with respect to $\theta$, we have that (31) is also twice-differentiable with respect to $\theta$. Thus, Proposition 10.1.6 in Cappé et al. (2005) applies for the SSR model, which completes the proof of Lemma 2.

Proof of Lemma 3. Define the conditional moments of the locally optimal transition density (46),

$$
\begin{equation*}
\mu_{t \mid t}^{x}:=\mathbb{E}_{\theta}\left[x_{t} \mid x_{t-1}, y_{t}\right] \quad \text { and } \quad \Sigma_{t \mid t}^{x}:=\operatorname{Var}_{\theta}\left[x_{t} \mid x_{t-1}, y_{t}\right] . \tag{A20}
\end{equation*}
$$

Applying the Gaussian projection, we can write these as

$$
\begin{align*}
\mu_{t \mid t}^{x} & =\mathbb{E}_{\lambda}\left[x_{t} \mid x_{t-1}\right]+\operatorname{Cov}_{\theta}\left[x_{t}, y_{t} \mid x_{t-1}\right] \operatorname{Var}_{\theta}\left[y_{t} \mid x_{t-1}\right]^{-1}\left(y_{t-1}-\mathbb{E}_{\theta}\left[y_{t} \mid x_{t-1}\right]\right) \\
& =\mu_{t \mid t-1}^{x}+\Sigma_{t \mid t-1}^{x} \Pi^{\prime}\left[\Sigma_{t \mid t-1}^{y}\right]^{-1}\left(y_{t}-\mu_{t \mid t-1}^{y}\right)  \tag{A21}\\
\Sigma_{t \mid t}^{x}= & \operatorname{Var}_{\lambda}\left[x_{t} \mid x_{t-1}\right]+\operatorname{Cov}_{\theta}\left[x_{t}, y_{t} \mid x_{t-1}\right] \Pi^{\prime} \operatorname{Var}_{\theta}\left[y_{t} \mid x_{t-1}\right]^{-1} \Pi \operatorname{Cov}_{\theta}\left[y_{t}, x_{t} \mid x_{t-1}\right] \\
= & \Sigma_{t \mid t-1}^{x}-\Sigma_{t \mid t-1}^{x} \Pi^{\prime}\left[\Sigma_{t \mid t-1}^{y}\right]^{-1} \Pi \Sigma_{t \mid t-1}^{x}, \tag{A22}
\end{align*}
$$

where we have used that,

$$
\begin{equation*}
\operatorname{Cov}_{\theta}\left[x_{t}, y_{t} \mid x_{t-1}\right]=\operatorname{Cov}_{\theta}\left[x_{t}, C\left(y_{0}\right)+\Pi x_{t}\left|x_{t-1}\right| x_{t-1}\right]=\Sigma_{t \mid t-1}^{x} \Pi^{\prime} \tag{A23}
\end{equation*}
$$

We define the conditional moments of the predictive observation density,

$$
\begin{align*}
\mu_{t \mid t-1}^{y} & :=\mathbb{E}_{\theta}\left[y_{t} \mid x_{t-1}\right]=\mathbb{E}_{\theta}\left[C\left(y_{0}\right)+\Pi x_{t} \mid x_{t-1}\right]=C\left(y_{0}\right)+\Pi \mu_{t \mid t-1}^{x}  \tag{A24}\\
\Sigma_{t \mid t-1}^{y} & :=\operatorname{Var}_{\theta}\left[y_{t} \mid x_{t-1}\right]=\operatorname{Var}_{\theta}\left[C\left(y_{0}\right)+\Pi x_{t} \mid x_{t-1}\right]=\Pi \Sigma_{t \mid t-1}^{x} \Pi^{\prime}+\Omega_{u} \tag{A25}
\end{align*}
$$

where we have used (22). Similarly, we define the conditional moments of the transition density,

$$
\begin{align*}
\mu_{t \mid t-1}^{x} & :=\mathbb{E}_{\lambda}\left[x_{t} \mid x_{t-1}\right]=\alpha+\Pi x_{t-1}  \tag{A26}\\
\Sigma_{t \mid t-1}^{x} & :=\operatorname{Var}_{\lambda}\left[x_{t} \mid x_{t-1}\right]=\Lambda_{t}, \tag{A27}
\end{align*}
$$

where we have used (23), which concludes the proof of Lemma 3.
Proof of Lemma 4. Lemma A. 4 establishes that Theorem 9.4.5 in Cappé et al. (2005) holds. It is a corollary to Theorem 9.4.5 in Cappé et al. (2005) that

$$
\begin{equation*}
\tilde{L}_{T}^{N}(\theta):=\prod_{t=1}^{T} W_{t}^{N} \quad \xrightarrow{P} \quad \boldsymbol{p}_{\theta}\left(y_{1: T}\right)=: \boldsymbol{L}_{T}(\theta), \tag{A28}
\end{equation*}
$$

as $N \rightarrow \infty$. By continuity of the logarithm, the continuous mapping theorem and the definitions (8) and (61), we therefore have that,

$$
\begin{equation*}
\tilde{\ell}_{T}^{N}(\theta)=\log \tilde{L}_{T}^{N}(\theta) \xrightarrow{P} \quad \log L_{T}(\theta)=\ell_{T}(\theta), \tag{A29}
\end{equation*}
$$

as $N \rightarrow \infty$, which completes the proof of Lemma A.4.
Proof of Lemma 5. We apply Lemma A. 5 for $t=T$ setting the test function to $\gamma_{T}\left(x_{1: T}\right):=U_{T}\left(x_{1: T} ; \theta\right)$, cf. (33). By Conjecture 2 we have that $U_{T}\left(x_{1: T} ; \theta\right) \in L^{2}\left[\mathcal{R}^{p \times T}, p_{\theta}\left(x_{1: T} \mid y_{1: T}\right)\right]$, which satisfies the condition, and Lemma A. 5 applies.

Proof of Lemma 6. We apply Lemma A. 4 to the functions $U_{T}\left(x_{1: T} ; \theta\right), V_{T}\left(x_{1: T} ; \theta\right)$ and the outer product $U_{T}\left(x_{1: T} ; \theta\right) U_{T}\left(x_{1: T} ; \theta\right)^{\prime}$ for $\theta \in \Theta$. First, Conjecture 2 implies that $U_{T}\left(x_{1: T} ; \theta\right) \in$ $L^{1}\left[\mathcal{R}^{p \times T}, \boldsymbol{p}_{\theta}\left(x_{1: T} \mid y_{1: T}\right)\right]$, such that by setting the test function to $\gamma_{T}\left(x_{1: T}\right):=U_{T}\left(x_{1: T} ; \theta\right)$, Lemma A. 4 gives us that,

$$
\begin{equation*}
\tilde{\mathbb{E}}_{\theta}^{N}\left[U_{T}\left(x_{1: T} ; \theta\right) \mid y_{1: t}\right] \quad \xrightarrow{P} \quad \mathbb{E}_{\theta}\left[U_{T}\left(x_{1: T} ; \theta\right) \mid y_{1: t}\right] \tag{A30}
\end{equation*}
$$

as $N \rightarrow \infty$. Second, Conjecture 2 states $V_{T}\left(x_{1: T} ; \theta\right) \in L^{1}\left[\mathcal{R}^{p \times T}, \boldsymbol{p}_{\theta}\left(x_{1: T} \mid y_{1: T}\right)\right]$, such that by setting the test function to $\gamma_{T}\left(x_{1: T}\right):=V_{T}\left(x_{1: T} ; \theta\right)$, Lemma A. 4 gives us that,

$$
\begin{equation*}
\tilde{\mathbb{E}}_{\theta}^{N}\left[V_{T}\left(x_{1: T} ; \theta\right) \mid y_{1: t}\right] \quad \xrightarrow{P} \quad \mathbb{E}_{\theta}\left[V_{T}\left(x_{1: T} ; \theta\right) \mid y_{1: t}\right], \tag{A31}
\end{equation*}
$$

as $N \rightarrow \infty$. Third, we note that by the Cauchy-Schwarz inequality it holds that,

$$
\begin{equation*}
\left\|U_{T}\left(x_{1: T} ; \theta\right) U_{T}\left(x_{1: T} ; \theta\right)^{\prime}\right\| \leq\left\|U_{T}\left(x_{1: T} ; \theta\right)\right\|\left\|U_{T}\left(x_{1: T} ; \theta\right)^{\prime}\right\|=\left\|U_{T}\left(x_{1: T} ; \theta\right)\right\|^{2} \tag{A32}
\end{equation*}
$$

such that, by Conjecture 2, we have that

$$
\begin{equation*}
\int\left\|U_{T}\left(x_{1: T} ; \theta\right) U_{T}\left(x_{1: T} ; \theta\right)^{\prime}\right\| \boldsymbol{p}_{\theta}\left(x_{1: T} \mid y_{1: T}\right) \mathrm{d} x_{1: T} \leq \int\left\|U_{T}\left(x_{1: T} ; \theta\right)\right\|^{2} \boldsymbol{p}_{\theta}\left(x_{1: T} \mid y_{1: T}\right) \mathrm{d} x_{1: T}<\infty \tag{A33}
\end{equation*}
$$

Thus, by setting the test function to $\gamma_{T}:=U_{T}\left(x_{1: T} ; \theta\right) U_{T}\left(x_{1: T} ; \theta\right)^{\prime}$, Lemma A. 4 gives us that,

$$
\begin{equation*}
\tilde{\mathbb{E}}_{\theta}^{N}\left[U_{T}\left(x_{1: T} ; \theta\right) U_{T}\left(x_{1: T} ; \theta\right)^{\prime} \mid y_{1: t}\right] \quad \xrightarrow{P} \quad \mathbb{E}_{\theta}\left[U_{T}\left(x_{1: T} ; \theta\right) U_{T}\left(x_{1: T} ; \theta\right)^{\prime} \mid y_{1: t}\right] \tag{A34}
\end{equation*}
$$

as $N \rightarrow \infty$. Now, by (37), (39), (40), (63), (66), and (67), we have that (A30)-(A34) correspond to,

$$
\begin{array}{rll}
\tilde{S}_{T}^{N}(\theta) & \xrightarrow{P} & \boldsymbol{S}_{T}(\theta) \\
\tilde{G}_{T}^{N}(\theta) & \xrightarrow{p} & \boldsymbol{G}_{T}(\theta) \\
\tilde{K}_{T}^{N}(\theta) & \xrightarrow{p} & \boldsymbol{K}_{T}(\theta), \tag{A37}
\end{array}
$$

as $N \rightarrow \infty$, respectively, such that we get by the continuous mapping theorem that,

$$
\begin{equation*}
\tilde{I}_{T}^{N}(\theta)=\tilde{S}_{T}^{N}(\theta) \tilde{S}_{T}^{N}(\theta)^{\prime}-\tilde{G}_{T}^{N}(\theta)-\tilde{K}_{T}^{N}(\theta) \quad \xrightarrow{P} \quad \boldsymbol{S}_{T}(\theta) \boldsymbol{S}_{T}(\theta)^{\prime}-\boldsymbol{K}_{T}(\theta)-\boldsymbol{G}_{T}(\theta)=\boldsymbol{I}_{T}(\theta), \tag{A38}
\end{equation*}
$$

as $N \rightarrow \infty$, which completes the proof of Lemma 6 .

## References

Anderson, Brian D. O., and John B. Moore. 1979. Optimal Filtering. Upper Saddle River: Prentice-Hall, pp. 1-367. [CrossRef]
Andrieu, Christophe, and Arnaud Doucet. 2002. Particle filtering for partially observed Gaussian state space models. Journal of the Royal Statistical Society B 64: 827-36. [CrossRef]
Bec, Frederique, and Anders Rahbek. 2004. Vector Equilibrium Correction Models with Non-Linear Discontinuous Adjustments. Econometrics Journal 7: 628-51. [CrossRef]
Bec, Frédérique, Anders Rahbek, and Neil Shephard. 2008. The ACR Model: A Multivariate Dynamic Mixture Autoregression. Oxford Bulletin of Economics and Statistics 70: 583-618. [CrossRef]
Cappé, Olivier, Eric Moulines, and Tobias Rydén. 2005. Inference in Hidden Markov Models. New York: Springer.
Chang, Yoosoon, J. Isaac Miller, and Joon Y. Park. 2009. Extracting a Common Stochastic Trend: Theory with Some Applications. Journal of Econometrics 150: 231-47. [CrossRef]
Chen, Rong, and Jun S. Liu. 2000. Mixture Kalman Filters. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62: 493-508. [CrossRef]
Chopin, Nicolas. 2004. Central Limit Theorem for Sequential Monte Carlo Methods and its Application to Bayesian Inference. The Annals of Statistics 32: 2385-411.
Creal, Drew. 2012. A Survey of Sequential Monte Carlo Methods for Economics and Finance. Econometric Reviews 31: 245-96.
Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 37: 1-38.
Dickey, David A., and Wayne A. Fuller. 1979. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. Journal of the American Statistical Association 74: 427-31. [CrossRef]
Dieci, Luca, and Erik S. Van Vleck. 1995. Computation of a Few Lyapunov Exponents for Continuous and Discrete Dynamical Systems. Applied Numerical Mathematics 17: 275-91.
Doornik, Jurgen A., and David F. Hendry. 2013. Modelling Dyamic Systems—PcGive 14: Volume II. London: Timberlake Consultants Ltd., pp. 1-284. [CrossRef]
Doornik, Jurgen A. 2012. Developer's Manual for Ox 7; London: Timberlake Consultants Ltd., pp. 1-182. [CrossRef]
Douc, Randal, Olivier Cappé, and Eric Moulines. 2005. Comparison of Resampling Schemes for Particle Filtering. Paper presented at the 4th International Symposium on Image and Signal Processing and Analysis, Zagreb, Croatia, September 15-17, pp. 64-69.
Douc, Randal, Eric Moulines, Jimmy Olsson, and Ramon Van Handel. 2011. Consistency of the Maximum Likelihood Estimator for General Hidden Markov Models. The Annals of Statistics 39: 474-513.
Douc, Randal, Eric Moulines, and David Stoffer. 2014. Nonlinear Time-Series: Theory, Methods, and Applications with $R$ Examples. Boca Raton: CRC Press. [CrossRef]
Doucet, Arnaud, and Neil Shephard. 2012. Robust Inference on Parameters via Particle Filters and Sandwich Covariance Matrices. Economics Series Working Papers 606, University of Oxford, Oxford, UK.
Doucet, Arnaud, Nando De Freitas, Kevin Murphy, and Stuart Russell. 2000. Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks. In Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc.

Doucet, Arnaud, Nando de Freitas, and Neil Gordon. 2001. Sequential Monte Carlo Methods in Practice. New York: Springer.
Durbin, James, and Siem Jan Koopman. 2012. Time Series Analysis by State Space Methods, 2nd ed. Oxford: Oxford University Press. [CrossRef]
Engle, Robert F., David F. Hendry, and Jean-Francois Richard. 1983. Exogeneity. Econometrica 51: 277-304. [CrossRef]
Engle, Robert F. 1982. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica 50: 987-1007. [CrossRef]
Feigin, Paul D., and Richard L. Tweedie. 1985. Random Coefficient Autoregressive Processes: A Markov Chain Analysis of Stationarity and Finiteness of Moments. Journal of Time Series Analysis 6: 1-14. [CrossRef]
Francq, Christian, and Jean-Michel Zakoian. 2010. GARCH Models: Structure, Statistical Inference and Financial Applications. Hoboken: Wiley, p. 489. [CrossRef]
Gordon, Neil J., David J. Salmond, and Adrian F. M. Smith. 1993. Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation. IEE Proceedings F, Radar and Signal Processing 140: 107-13. [CrossRef]
Granger, Clive W. J., and Norman R. Swanson. 1997. An Introduction to Stochastic Unit-Root Processes. Journal of Econometrics 80: 35-62. [CrossRef]
Hamilton, James Douglas. 1994. Time Series Analysis. Princeton: Princeton University Press.
Jensen, Søren Tolver, and Anders Rahbek. 2004. Asymptotic Inference for Nonstationary GARCH. Econometric Theory 20: 1203-26. [CrossRef]
Johansen, Søren. 1996. Likelihood-Based Inference in Cointegrated Vector Autoregressive Models (Advanced Texts in Econometrics). Oxford: Oxford University Press.
Juselius, Katarina. 2007. The Cointegrated VAR Model: Methodology and Applications (Advanced Texts in Econometrics). Oxford: Oxford University Press. [CrossRef]
Kalman, Rudolph Emil. 1960. A New Approach to Linear Filtering and Prediction Problems. Transactions of the ASME-Journal of Basic Engineering 82: 35-45.
Kristensen, Dennis, and Anders Rahbek. 2010. Likelihood-based Inference for Cointegration with Nonlinear Error-Correction. Journal of Econometrics 158: 78-94.
Kristensen, Dennis, and Anders Rahbek. 2013. Testing and Inference in Nonlinear Cointegrating Vector Error Correction Models. Econometric Theory 29: 1238-88. [CrossRef]
Kushner, Harold, and G. George Yin. 2003. Stochastic Approximation and Recursive Algorithms and Applications, 2nd ed. New York: Springer.
Leybourne, Stephen J., Brendan P. M. McCabe, and Andrew R. Tremayne. 1996. Can Economic Time Series Be Differenced To Stationarity? Journal of Business \& Economic Statistics 14: 435-46.
Lieberman, Offer, and Peter C. B. Phillips. 2014. Norming rates and limit theory for some time-varying coefficient autoregressions. Journal of Time Series Analysis 35: 592-623.
Lieberman, Offer, and Peter C. B. Phillips. 2017. A multivariate stochastic unit root model with an application to derivative pricing. Journal of Econometrics 196: 99-110. [CrossRef]
Ling, Shiqing. 2004. Estimation and Testing Stationarity for Double-Autoregressive Models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66: 63-78. [CrossRef]
Ling, Shiqing. 2007. A Double AR(p) Model: Structure and Estimation. Statistica Sinica 17: 161-75.
Louis, Thomas A. 1982. Finding the Observed Information Matrix when Using the EM Algorithm. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 44: 226-33. [CrossRef]
McCabe, Brendan P. M., and Richard J. Smith. 1998. The power of some tests for difference stationarity under local heteroscedastic integration. Journal of the American Statistical Association 93: 751-61.
McCabe, Brendan P. M., and Andrew R. Tremayne. 1995. Testing a time series for difference stationarity. Annals of Statistics 23: 1015-28.
Meyn, Sean P., and Richard L. Tweedie. 2005. Markov Chains and Stochastic Stability; Cambridge: Cambridge University Press. [CrossRef]
Murray, Iain, Zoubin Ghahramani, and David MacKay. 2006. MCMC for Doubly-Intractable Distributions. Paper presented at the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI2006), Cambridge, MA, USA, July 13-16.
Nocedal, Jorge, and Stephen Wright. 2006. Numerical Optimization, 2nd ed. New York: Springer. [CrossRef]

Olsson, Jimmy, and Tobias Rydén. 2008. Asymptotic Properties of Particle Filter-Based Maximum Likelihood Estimators for State Space Models. Stochastic Processes and Their Applications 118: 649-80.
Pedersen, Rasmus Søndergaard, and Olivier Wintenberger. 2018. On the tail behavior of a class of multivariate conditionally heteroskedastic processes. Extremes 21: 261-84.
Pitt, Mark, and Neil Shephard. 1999a. Time Varying Covariances: A Factor Stochastic Volatility Approach. In Bayesian Statistics. Edited by José M. Bernardo, James O. Berger, A. P. Dawid and Adrian F. M. Smith. Oxford: Oxford University Press, vol. 6, pp. 547-70.
Pitt, Michael K., and Neil Shephard. 1999b. Filtering via Simulation: Auxiliary Particle Filters. Journal of the American Statistical Association 94: 590-99. [CrossRef]
Polyak, Boris T., and Anatoli B. Juditsky. 1992. Acceleration of Stochastic Approximation by Averaging. SIAM Journal on Control and Optimization 30: 838-55.
Polyak, Boris Teodorovich. 1990. A New Method of Stochastic Approximation Type. Automation and Remote Control 51: 98-107. [CrossRef]
Poyiadjis, George, Arnaud Doucet, and Sumeetpal S. Singh. 2011. Particle Approximations of the Score and Observed Information Matrix in State Space Models with Application to Parameter Estimation. Biometrika 98: 65-80.
Robbins, Herbert, and Sutton Monro. 1951. A Stochastic Approximation Method. The Annals of Mathematical Statistics 22: 400-7. [CrossRef]
Robert, Christian P., and George Casella. 2010. Introducing Monte Carlo Methods with R, 1st ed. New York: Springer, p. 296. [CrossRef]

Nielsen, Heino Bohn, and Anders Rahbek. 2014. Unit Root Vector Autoregression with Volatility Induced Stationarity. Journal of Empirical Finance 29: 144-67. [CrossRef]
Nielsen, Heino Bohn. 2016. The Co-integrated Vector Autoregression with Errors-in-Variables. Econometric Reviews 35: 169-200. [CrossRef]
© 2018 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http:/ / creativecommons.org/licenses/by/4.0/).

# Article <br> A Parametric Factor Model of the Term Structure of Mortality 

Niels Haldrup and Carsten P. T. Rosenskjold *<br>Center for Research in Econometric Analysis of Time Series (CREATES), Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, DK-8210 Aarhus V, Denmark; nhaldrup@econ.au.dk<br>* Correspondence: cthillemann@econ.au.dk

Received: 11 June 2018; Accepted: 5 March 2019; Published: 11 March 2019


#### Abstract

The prototypical Lee-Carter mortality model is characterized by a single common time factor that loads differently across age groups. In this paper, we propose a parametric factor model for the term structure of mortality where multiple factors are designed to influence the age groups differently via parametric loading functions. We identify four different factors: a factor common for all age groups, factors for infant and adult mortality, and a factor for the "accident hump" that primarily affects mortality of relatively young adults and late teenagers. Since the factors are identified via restrictions on the loading functions, the factors are not designed to be orthogonal but can be dependent and can possibly cointegrate when the factors have unit roots. We suggest two estimation procedures similar to the estimation of the dynamic Nelson-Siegel term structure model. First, a two-step nonlinear least squares procedure based on cross-section regressions together with a separate model to estimate the dynamics of the factors. Second, we suggest a fully specified model estimated by maximum likelihood via the Kalman filter recursions after the model is put on state space form. We demonstrate the methodology for US and French mortality data. We find that the model provides a good fit of the relevant factors and, in a forecast comparison with a range of benchmark models, it is found that, especially for longer horizons, variants of the parametric factor model have excellent forecast performance.


Keywords: mortality forecasting; term structure of mortality; factor modelling; cointegration
JEL Classification: C1; C22; J10; J11; G22

## 1. Introduction

The Lee and Carter (1992) (LC) model has become a benchmark model when estimating and forecasting improvements in age-specific death rates and the calculation of life expectancy. The model is basically a one-factor model that allows for a single common time trend with age-specific loadings. The model has been extended in many different ways. For instance, Booth et al. (2002) and Renshaw and Haberman (2003) extend the model with a second common time trend with age-specific loadings. Hyndman and Ullah (2007) developed a functional data approach in which the data are smoothed across age prior to modelling using penalized regression splines and principal component analysis. We will refer to these models as nonparametric factor models. De Jong and Tickle (2006) use the state space framework to establish smoothness in the LC model using b-splines.

Typically, the estimation of factors is implemented nonparametrically via either singular value decomposition or principal component analysis. For models with multiple factors, these are identified via orthogonalization. Subsequently, the factors are modelled as individual time series models which can be used for forecast projections.

The LC model and its extensions are basically statistical models that summarize the variability of the measured age-specific death rates over time in a parsimonious way. No structure is imposed in the
model specification. However, in the demographics literature on mortality laws, it is well established that age groups are exposed rather differently to death risk and it seems reasonable to believe that separate time factors may affect different age groups rather than assuming a single common factor as in the basic LC model.

Mortality laws for death rates observed at a given time have been suggested by amongst others Gompertz (1825); Makeham (1860); and Heligman and Pollard (1980); Tabeau et al. (2001) provide a review. These laws refer to separate mortality characteristics for different age groups such as infants, youths, adults, and the elderly. When accounting for the dynamic development of mortality over time, it seems natural to consider a factor model that accounts for these mortality laws. In this paper, we assume the presence of multiple factors and impose structure on the loadings via specific functional forms. The approach is similar to McNown and Rogers (1989). However, their model is both heavily over parametrized in terms of latent time-varying parameters and does not fully exploit the information contained in the factor dynamics of the model.

The idea is similar to e.g., the dynamic Nelson-Siegel model for the term structure of interest rates-see Diebold and Li (2006). Diebold and Li suggest a factor model with parametrized factor loadings which identify level, slope, and curvature of the yield curve, associated with the long, short, and medium term yields. In the context of the term structure of mortality, we define loading functions that identify the factors that drive infant, adult, and 'accident hump' (youth) mortality, respectively, plus a common factor uniformly affecting all age groups. We will generally refer to this class of factor models as parametric factor models (PFM). It follows from this approach that, opposed to traditional factor analysis, the factors to be extracted will not necessarily be independent. In fact, the factors may potentially cointegrate when these are found to have unit roots.

We consider estimation of the model parameters and the factors by cross-section regressions over age groups for each period of time. These estimations are conducted over a grid of tuning parameters that define the shape of the loading functions. Next, a least squares criterion is used to determine the desired tuning parameters and the corresponding factor elements. This approach is similar to the first step of the cross section projection procedure suggested in Diebold and Li (2006). After the factors have been extracted, the second step implies the estimation of a time series model for the factors. This can be done in different ways. For instance, univariate as well as multivariate models for the factors can be formulated and with the possibility of stationary as well as non-stationary factors that potentially cointegrate. The final time series specification of the factor dynamics is an empirical question and separate time series models are considered for this purpose.

We also consider a fully parametrized model specification formulated as a state space model. By use of the Kalman filter recursions, the model parameters and the factors can be estimated by full maximum likelihood. This approach is similar to that of Diebold et al. (2006) for the term structure of interest rates.

The proposed model for women and men is estimated using French and US data for the sample period 1950-2014. The estimated functional forms appear to be rather similar across countries with the duration of the accident hump being longer for men than for women. The shape of the four factors also generally appear similar across countries but with differences across genders. In terms of model fit compared with the LC model, it appears that our model is doing especially well for explaining the age-specific death rates around the age groups defining the 'accident hump'.

We also evaluate the out of sample performance of the model where the predicted mortality rates are summarized in a loss function defined by the life expectancy. Specifically, we apply the model confidence set procedure of Hansen et al. (2011) to evaluate the relative forecast performance on the horizons of 1,10 , and 20 years ahead using a number of benchmark models. We find that, particularly for long horizon forecasts, our model tends to be in the set of best predicting models.

In Section 2, we briefly describe the LC model and provide a detailed description of the mortality data and set up a number of stylized facts of the mortality curve in Section 3. Section 4 introduces the PFM and its interpretation. Section 5 discusses estimation of the PFM and in Section 6 we present the
empirical analysis, including the estimation results, the model fit, and the factor dynamics. Section 7 examines the relative out-of-sample forecast performance. Section 8 provides conclusions.

## 2. The Lee-Carter Model

The observed data of the analysis are the age specific death rates $m_{x, t}$ for age groups $x=0,1, \ldots, N$ at year $t=t_{0}, \ldots, T$, broadly defined as the number of deaths at age $x$ in year $t$ divided by the Exposure-to-Risk, which is the average population aged $x$ in year $t$. The data used is obtained from the Human Mortality Database (2016).

The Lee and Carter (1992) (LC) model describes the (log) age-specific death rates by:

$$
\begin{equation*}
\ln m_{x, t}=\alpha_{x}+\beta_{x} \kappa_{t}+\varepsilon_{x, t}, \tag{1}
\end{equation*}
$$

where $\alpha_{x}$ captures the average death rate for each age $x . \kappa_{t}$ is a common time varying factor capturing the general trend in death rates over time $t . \beta_{x}$ is the factor loading capturing the effect of the factor $\kappa_{t}$ on each age group $x$ and $\varepsilon_{x, t}$ is the age and time specific error term. The LC model is basically a one-factor model that allows a common time trend to have age-specific loadings with respect to the development of the age-specific $\log$ death rates. Lee and Carter (1992) obtain identification using the normalizations $\sum_{x=0}^{N} \beta_{x}=1$ and $\sum_{t=t_{0}}^{T} \kappa_{t}=0$. The constraints imply that $\alpha_{x}$ measures the age specific time-average of the $\log$ death rates, $\ln m_{x, t} .{ }^{1}$ To estimate $\beta_{x}$ and $\kappa_{t}$, the singular value decomposition is applied to the matrix $(A)_{x t}=\ln m_{x t}-\alpha_{x}$ for all $x, t$. Lee and Carter (1992) find that $\kappa_{t}$ can be modelled as a random walk with drift, although they allow for other specifications as well. The LC model is designed to maximize the in-sample fit by fitting a general factor model structure to the death rates. Note that the LC model does not impose any particular structure on the age-specific graduation of mortality, which essentially is data driven. However, it imposes a rigid structure on the improvements of the age-specific death rates over time by requiring these to be proportional and governed by the single factor $\mathcal{K}_{t}$.

## 3. Stylized Facts of the Mortality Curve

A good mortality model should desirably account for both the age (cross section) dimension of mortality as well as its development over time, i.e., the time dimension. Here, we describe some stylized facts of the (log) death rates to be modelled. For illustration, we use data for France and USA available from the Human Mortality Database (2016). ${ }^{2}$

The age dimension: To illustrate the age dimension properties, which a good mortality model should be able to capture, we show the log mortality on 10 year intervals from 1950 to 2010 for men and women in Figure 1a-d for the US and France. The mortality curve shows a similar shape over the ages, but the level of mortality tends to decline over time; the shape is very similar across both genders and countries. The infant mortality is seen to decline rapidly during early childhood. In the late teens, the mortality rate experiences a rapid increase often termed the 'accident hump', which appears either as a distinct hump or as a flattening out of the death rates; see Heligman and Pollard (1980). After the accident hump, the mortality rates are gradually increasing with age (log-linearly). Thus, for a model to produce realistic results, these three facts should hold for each year. The three properties could also be interpreted in terms of biological reasonableness as described by Cairns et al. (2006), which rules out patterns that are biologically unreasonable such as a decreasing mortality curve for the older as well as the crossing over of age-specific mortality rates.

[^34]

Figure 1. The log age-specific death rates for the years 1950, 1960, 1970, 1980, 1990, 2000, and 2010 for men and women in France and the USA.

The time dimension: When investigating the time dynamics in the development of the log age-specific death rates, the review paper by Wong-Fupuy and Haberman (2004, p. 56) notes that "There is a broad consensus across the resulting projections: (1) an approximately log-linear relationship between mortality rates and time, (2) decreasing improvements according to age"' The first point helps to explain the success of the LC model where the common time-varying factor is found to evolve almost linearly in most applications-see, e.g., Lee and Miller (2001) and Callot et al. (2016). The log-linear development of death rates over time is illustrated in Figure 2a-d. The second observation of decreasing improvements in mortality with respect to age can be described by the so called compensation effect of mortality—see, e.g., Gavrilov and Gavrilova (1979, 1991). ${ }^{3}$ In Figure 2a-d, this effect is seen by a slope of the $\log$ mortality-time plot that decreases with age.

Several studies find that a unit root is present in the individual age-specific death rates-see, for instance, Lazar and Denuit (2009). In addition, it is common that the time-factor of the LC model is modelled as a random walk with drift. Basically, this means that all death rates are governed by the same stochastic time trend component and hence for a system of $N+1$ age groups, all death rates cointegrate pairwisely and a total of $N$ cointegrating relations exists among all age-specific death rates.

[^35]

Figure 2. The log age-specific death rates for a range of ages for French and American men and women from 1950-2014.

To examine this feature of the LC model, we have conducted cointegration tests of all the pairwise combinations of log mortality across age groups. Note that, due to the dimension of the data, a full cointegration analysis cannot be conducted for the full data set. Figure 3a,d report a heatmap of the $p$-values from Johansen's trace test, Johansen (1991), of the null hypothesis of zero cointegrating relations against one cointegrating relation for all combinations of the log age-specific death rates. As seen, we cannot reject the null of no cointegration for most of the death rate pairs, especially for US data. It is apparent that most of the cointegrating relations found are between the adjacent ages, i.e., along the diagonal line. For both countries, we find clear rejection of no cointegration amongst the youngest ages, but not for newborns. For France, rejection of no cointegration among the oldest ages is found to a larger degree. Furthermore, it is found that around the accident hump and for the infants we cannot reject the null for relatively adjacent ages. Thus, overall the figures clearly show that the assumption of the LC model of $N$ cointegrating relations is not consistent with the mortality data. We note that this is also consistent with Lazar and Denuit (2009) who found multiple stochastic trends when investigating cointegration across seven age groups of five-year age intervals. The stochastic trends driving mortality over time are generally different across the age groups. The model we are subsequently going to propose will not have the restrictive feature of the LC model, since different factors are constructed to affect separate age groups.

In summary, we observe seven stylized facts for the term structure of mortality that a good mortality model should be able to reproduce: (1) declining mortality for infants, (2) increasing mortality around the accident hump, (3) log-linearly increasing mortality with age for adults, (4) a log-linear relationship between the death rates and time, (5) the log age-specific death rates are integrated of order one around a linear trend, (6) decreasing improvements in mortality with age, and (7) multiple stochastic trends characterize the development of log mortality over time for the different age groups.


Figure 3. $p$-Values from Johansen's Trace test are shown for all pairwise combinations of the (log) age-specific death rates for French and US men and women over the period 1950-2014. The test is performed with a restricted time trend in the cointegrating relation and 1 lag in the VAR specification. The $p$-Values are obtained via the gamma approximation following Doornik $(1998,1999)$ and shown for significance levels between 0 and 0.10 .

## 4. The Parametric Factor Model for the Term Structure of Mortality

The model we propose assumes that mortality is driven by multiple factors and we impose structure on the factor loadings capturing the regularities discussed in the previous section.

The PFM reads as follows:

$$
\begin{equation*}
\ln m_{x, t}=\kappa_{0, t}+\kappa_{1, t} e^{-\lambda_{1} x}+\kappa_{2, t} e^{-\lambda_{2}(\ln (x)-\ln (k))^{2}}+\kappa_{3, t}\left(\frac{x}{N}\right)^{\lambda_{3}}+\varepsilon_{x, t} . \tag{2}
\end{equation*}
$$

The model has four factors $\kappa_{i, t}, i=0,1,2,3$ with loading functions that are designed to capture distinct age groups. The shape of the loading functions are governed by the constant parameters $\lambda_{1}, \lambda_{2}, \lambda_{3}$ and $k$ which are assumed positive. $N$ is the maximum age used for the analysis, which is set to 95 due to data quality, as described in Section 3. $\kappa_{0, t}$ is a factor that is common to all age groups. The factor $\kappa_{1, t}$ captures child mortality, $\kappa_{2, t}$, the accident hump, and finally, $\kappa_{3, t}$ is a factor that tends to increase mortality with age. Note that the common factor has the constant loading one for all age groups. The loading for infant mortality declines rapidly with age. The loading for the accident hump is approximately bell-shaped around age $k$, which is estimated to be the age at which the accident hump equals one-see Figure $4 \mathrm{a}, \mathrm{b}$ below. Finally, the loading for the adult factor grows almost linearly with age when $\lambda_{3}$ is close to one. The error term $\varepsilon_{x, t}$ is assumed to be IID normally distributed as $N\left(0, \sigma^{2}\right)$ for all ages and years. ${ }^{4}$ The loading functions estimated for France and the US are shown in Figure $4 \mathrm{a}, \mathrm{b}$; the estimation procedure will be discussed in the next section. Even though it may be claimed that the functional forms of the loading functions are arbitrary, they are designed such that the mortality laws and stylized facts, described in Section 3, are captured through the model specification.

The level and infant terms $\kappa_{0, t}$ and $\kappa_{1, t} e^{-\lambda_{1} x}$, respectively, are used in many models using the age-specific graduation of mortality, see e.g., Siler (1979) and Rogers and Little (1994). The accident hump loading $e^{-\lambda_{2}(\ln (x)-\ln (k))^{2}}$ is taken from Heligman and Pollard (1980). The adult factor can be seen as a generalization of the Gompertz model, inspired by the Box and Cox (1964) power formulation. That is, the loading function captures the Gompertz specification if $\lambda_{3}=1$.

It is clear that the single factors $\kappa_{i, t}$ are only identified when $\lambda_{i}$ is non-zero. If some $\lambda_{i}$ is zero, it means that the associated factor is absent and can be left out from the analysis. Identification of factors when $\lambda_{i}$ are non-zero is a result of imposing a particular functional form on the loadings. Hence, the identification issue of the Lee-Carter model is absent in the present model. See Nielsen and Nielsen (2014) about a general discussion of identification in mortality models.


Figure 4. Cont.

[^36]
(b) Loadings, USA.

Figure 4. Plot of the estimated loading functions for the years 1950-2014 for men and women in France and USA. The loading functions correspond to the level, infant, accident hump, and adult age groups, respectively. The loadings are estimated following the two-step procedure described in Section 5.

## 5. Estimation Procedure for the Parametric Factor Model

We consider two estimation procedures for estimating the PFM, the two-step procedure of Diebold and Li (2006) and exact maximum likelihood estimation using the Kalman Filter recursions of the model written on state-space form similar to Diebold et al. (2006). Alternatively, one could use maximum likelihood estimation following Brouhns et al. (2002) assuming a Poisson distribution for the death counts.

### 5.1. The Two-Step Estimation Procedure

The two-step procedure considers first to estimate the model parameters and the factors of the model and, second, estimating a time series model of the extracted factors with the primary purpose of forecasting. Regarding the first step, McNown and Rogers (1989) propose to estimate the factors by nonlinear least squares for each point in time, hence giving a time series of the factors. This allows not only the factors but also the model (loading) parameters to be time-varying. McNown and Rogers (1992) fix the parameters of the model a priori and estimate the factors in a sequence of cross-section regressions. The latter procedure is also the one adopted by Diebold and Li (2006) when estimating the dynamic Nelson-Siegel model for the term structure of interest rates, where the different loadings refer to the level, slope, and curvature of the yield curve.

We suggest modifying McNown and Rogers (1992) and Diebold and Li (2006) by considering cross-sectional regressions at each time $t$ for a fine grid of the model parameters and select the preferred model by minimizing the conditional sum of squares function. Hence, for a given set of loading parameters, the factors can simply be estimated by using ordinary least squares for each year. This can also be implemented by a nonlinear least squares optimization algorithm. Here, we use the limited memory BFGS procedure ("L-BFGS-B") developed by Byrd et al. (1995) and implemented via the R package 'Optim' (R Core Team 2015). This step provides estimates of the four factors of the model. Note that, as opposed to traditional factor models, generally the estimated factors will not be orthogonal and in fact are most likely to be dependent. In the second step of the two-step procedure, time-series models are fitted to the factors. This step is only needed when the model is used
for predictions as we shall see in Section 7. These can be univariate time series models such as ARIMA model specifications, possibly with drifts or trends, or the factors can be modelled as stationary or nonstationary VAR models which potentially can allow for cointegration amongst the factors. It is an empirical question to properly select a time series model in the second step.

### 5.2. One-Step Estimation

The parametric factor model in Equation (2) can be formulated on state-space form and estimated by maximum likelihood by use of the Kalman Filter, see e.g., Durbin and Koopman (2012). This estimation procedure improves on the two-step estimation procedure by allowing joint estimation of the latent factors and their transition dynamics as well as the unknown parameters assuming Gaussian errors. Estimating the system jointly delivers the appropriate likelihood quantities, unlike the two-step approach, which ignores the uncertainty and estimation errors from the first step. We conjecture that standard inference results hold for the Gaussian Maximum Likelihood approach, although we do not provide a formal proof for this; see also Koopman et al. (2010).

The measurement equation of the state space model can be written as:

$$
\ln \mathbf{m}_{t}=\Lambda \kappa_{t}+\varepsilon_{t}
$$

where

$$
\ln \mathbf{m}_{t}=\left(\begin{array}{c}
\ln m_{0, t} \\
\ln m_{1, t} \\
\vdots \\
\ln m_{N, t}
\end{array}\right), \varepsilon_{t}=\left(\begin{array}{c}
\varepsilon_{0, t} \\
\varepsilon_{1, t} \\
\vdots \\
\varepsilon_{N, t}
\end{array}\right), \kappa_{t}=\left(\begin{array}{c}
\kappa_{0, t} \\
\kappa_{1, t} \\
\kappa_{2, t} \\
\kappa_{3, t}
\end{array}\right)
$$

and

$$
\Lambda=\left(\begin{array}{cccc}
1 & e^{-\lambda_{1} \cdot 0} & e^{-\lambda_{2}(\ln (0)-\ln (k))^{2}} & \left(\frac{0}{N}\right)^{\lambda_{3}} \\
1 & e^{-\lambda_{1} \cdot 1} & e^{-\lambda_{2}(\ln (1)-\ln (k))^{2}} & \left(\frac{1}{N}\right)^{\lambda_{3}} \\
\vdots & \vdots & \vdots & \vdots \\
1 & e^{-\lambda_{1} \cdot N} & e^{-\lambda_{2}(\ln (N)-\ln (k))^{2}} & \left(\frac{N}{N}\right)^{\lambda_{3}}
\end{array}\right)
$$

As in Section 4, the vector error term $\varepsilon_{t}$ is assumed to be normally distributed as $N\left(0, I \sigma^{2}\right)$, where $I$ is the identity matrix.

The transition equation of the state space model should be formulated to capture the dynamics of the factors. For instance, if we assume that the factors are governed by a VAR(1) process in first differences, the transition equation can be specified as:

$$
\binom{\kappa_{t}}{\Delta \kappa_{t+1}}=\left[\begin{array}{cc}
I_{4} & I_{4} \\
0 & \Phi
\end{array}\right]\binom{\kappa_{t-1}}{\Delta \kappa_{t}}+\binom{0}{c}+\binom{0}{v_{t}}
$$

where $v_{t}$ is multivariate normal distributed as $N(0, \Sigma)$ and $c$ is a four-dimensional vector constant.
In the case where the factors cointegrate with $r$ cointegrating relations, the transition dynamics can be written as:

$$
\binom{\kappa_{t}}{\Delta \kappa_{t+1}}=\left[\begin{array}{cc}
I_{4} & I_{4} \\
\alpha \gamma^{\prime} & \alpha \gamma^{\prime}
\end{array}\right]\binom{\kappa_{t-1}}{\Delta \kappa_{t}}+\binom{0}{c}+\binom{0}{v_{t}}
$$

where $v_{t}$ is multivariate normal distributed as $N(0, \Sigma)$. Both $\alpha$ and $\gamma$ are $4 \times r$ matrices. The second row gives the desired VECM specification for the transition dynamics:

$$
\begin{equation*}
\Delta \kappa_{t+1}=\alpha \gamma^{\prime} \kappa_{t}+c+v_{t} . \tag{3}
\end{equation*}
$$

Note that the constant $c$ is treated as a state parameter vector within the Kalman filter. Estimation of the parameters $\psi=\left[\lambda_{1}, \lambda_{2}, \lambda_{3}, k, \sigma, \Sigma, \Phi(\right.$ or $\left.\alpha, \gamma), c\right]$ is achieved via numerical optimization of the prediction error decomposition of the likelihood function:

$$
\begin{equation*}
\mathcal{L}(\psi)=-\frac{N T}{2} \ln 2 \pi-\frac{1}{2} \sum_{t=1}^{T} \ln \left|F_{t}\right|-\frac{1}{2} \sum_{t=1}^{T} v_{t}^{\prime} F_{t}^{-1} v_{t} \tag{4}
\end{equation*}
$$

where $v_{t}$ is the one step (innovation) prediction error of the measurements equation and $F_{t}$ is the innovation covariance matrix of the measurement equation. The numerical optimization is performed via the low-memory BFGS procedure "L-BFGS-B" from Byrd et al. (1995) in the R package Optim (R Core Team 2015).

## 6. Empirical Analysis

### 6.1. Estimates Using the Two-Step Procedure

Figure 4a,b in Section 4 display the estimated shape of the loading functions for French and US men and women based on the two-step procedure. Table 1 reports the estimated shape parameters and their standard errors. Note that all parameters are significantly different from zero and hence the factors are identified. The estimated parameters are similar across countries. However, the loading functions for the adult curvature are more convex for women than for men. Similarly, the shape and location of the accident hump vary across genders with men suffering from the accident hump longer than for women.

Table 1. Estimated loading function parameters and standard errors from the first step in the two-step procedure for French and US men and women. The standard errors are calculated using the inverse Fisher information criterion.


The estimated factors are shown in Figure 5a-d for France and Figure 6a-d for the US.
A number of insights follow from these figures. The factor governing the common mortality level decreases almost linearly and thus capturing a common decline in mortality across all age groups; this applies for both genders and countries. The infant factor for both men and women decline over the period showing that the infants have seen larger improvements in mortality reduction compared to the general level captured by the first factor. Moreover, it can be seen that the decline for the infant factor stagnates around 1995 for all populations considered. Hence, after 1995, the development in mortality for infants has generally followed the common rate.

The accident hump factor shows an increase in size from 1950 to about 1990 followed by stagnation for all but US men. Regarding the development of the adult factor, Figures 5d and 6d exhibit an upward slope over the sample period and hence reduce the mortality improvements for the relevant age group.

Thus, slower improvements in mortality with age are captured by the model, in line with the stylized facts previously reported.


Figure 5. The factors $\kappa_{i, t}, i=0,1,2,3$ estimated by the two-step procedure for France using data from 1950-2014. The plots are showing the level factor, infant factor, accident hump factor, and adult factor for both genders, respectively.


Figure 6. The factors $\kappa_{i, t}, i=0,1,2,3$ estimated by the two-step procedure for USA using data from 1950-2014. The plots are showing the level factor, infant factor, accident hump factor, and adult factor for both genders, respectively.

### 6.2. Cointegrating Analysis of the Factors

In order to use the estimated model for forecast projections, we need to examine the time series features of the estimated factors $\kappa_{i, t}, i=0,1,2,3$. By using a range of unit root tests, we find strong empirical support for the presence of unit roots, possibly with a drift, in all of the factors considered across both countries and gender. Given this observation, it is not surprising that the age-specific $\log$ death rates individually appear to have similar time series characteristics. The one-factor model of Lee and Carter (1992) also typically model the factor as a random walk with drift. From visual inspection of the factors in Figures 5 and 6, it is evident that the various factors tend to co-move across gender and thus the factors are likely to cointegrate. Accounting for cointegration amongst the factors will potentially lead to superior forecasts.

We have conducted cointegration analysis using the Johansen (1988) trace test for different subsets of factors. In Table 2, we report the test results for each country and for each gender using all four factors. In Table 3, we examine tests for each country using all eight factors for both men and women, and, finally, Table 4 displays the tests for men and women, respectively, by merging the factors across countries.

The results are rather different for the USA and France as can be seen from Table 2. For both genders, the US factors are found not to cointegrate and hence these factors are driven by four separate common stochastic trends. On the other hand, the factors for French men and women appear to cointegrate with two or three cointegrating vectors and thus the factors for each gender appear to be driven by a single or possibly two common stochastic trends. This finding is also in line with the heat maps reported in Figure 3 showing that, for France, the pairwise log mortality rates appear more cointegrated compared to the USA.

Table 2. Test for cointegration rank amongst factors for US and French men and women.

| USA |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
| Men |  |  |  |  |
| Rank | Trace-Test | $p$-Value | Trace-Test | $p$-Value |
| 0 | 52.240 | $[0.323]$ | 47.939 | $[0.512]$ |
| 1 | 27.748 | $[0.641]$ | 30.689 | $[0.468]$ |
| 2 | 13.926 | $[0.668]$ | 15.243 | $[0.562]$ |
| 3 | 1.1522 | $[0.992]$ | 3.0677 | $[0.858]$ |
| France |  |  |  |  |
| 0 | 106.790 | $[0.000]^{* *}$ | 108.730 | $[0.000]^{* *}$ |
| 1 | 56.044 | $[0.001]^{* *}$ | 47.599 | $[0.014]^{*}$ |
| 2 | 28.089 | $[0.024]^{*}$ | 24.893 | $[0.064]$ |
| 3 | 3.642 | $[[0.788]$ | 9.2236 | $[0.171]$ |

Note: The Johansen trace test is calculated with a trend restricted to the cointegration space. The number of lags in the VAR is 1 for all cases. "**" and "**" signify significance at the $1 \%$ and $5 \%$ level, respectively.

In Table 3, the set of variables in the VAR model is expanded to include both men and women for each country. In this case, the eight factors for the US data are driven by four common stochastic trends. It is tempting to believe that the factors cointegrate across genders, however, a formal statistical test rejects this hypothesis. For the French data, the eight factors have six to seven cointegrating vectors and thus have one or two common stochastic trends. Again, a formal test rejects that the factors cointegrate pairwisely across genders.

Finally, Table 4 shows that, when pooling the US and French data for men and women, respectively, both the male and female factors are likely to be driven by six factors and thus have two cointegrating relations. Hence, cross-country similarities exist across countries for both genders but only to a limited extent.

These findings demonstrate that different time series specifications should be considered when modelling the factors with the purpose of forecasting. For the US, it seems appropriate to specify a VAR in first differences with a vector of unrestricted constants to capture the drift of the single series. It could also be considered to base predictions on an expanded (cointegrated) VAR model including factors for both genders. For France, a cointegrated VAR with cointegration rank two or three seems appropriate. An expanded cointegrated VAR with eight factors and six to seven cointegrating vectors is also possible. When modelling the factors as univariate time series models, a random walk with drift specification is appropriate, but, since the cross dependence of factors is neglected in this case, it is likely that inferior forecasts will result.

Table 3. Test for cointegration rank amongst factors for men and women for USA and France.

|  | USA |  | France |  |
| :---: | :--- | :--- | :--- | :--- |
|  | Men and Women |  | Men and Women |  |
| Rank | Trace-Test | $p$-Value | Trace-Test | $p$-Value |
| 0 | 286.700 | $[0.000]^{* *}$ | 287.730 | $[0.000]^{* *}$ |
| 1 | 194.530 | $[0.000]^{* *}$ | 215.400 | $[0.000]^{* *}$ |
| 2 | 139.460 | $[0.001]^{* *}$ | 158.870 | $[0.000]^{* *}$ |
| 3 | 89.751 | $[0.041]^{*}$ | 116.050 | $[0.000]^{* *}$ |
| 4 | 52.267 | $[0.321]$ | 76.354 | $[0.002]^{* *}$ |
| 5 | 34.795 | $[0.257]$ | 47.410 | $[0.015]^{*}$ |
| 6 | 19.806 | $[0.240]$ | 24.016 | $[0.082]$ |
| 7 | 7.897 | $[0.268]$ | 8.524 | $[0.218]$ |

Note: The Johansen trace test is calculated with a trend restricted to the cointegration space. The number of lags in the VAR is 1 for all cases. "**" and "*" signify significance at the $1 \%$ and $5 \%$ level, respectively.

Table 4. Test for cointegration rank amongst factors for US and French men and women.

|  | Men |  | Women |  |
| :---: | :--- | :--- | :--- | :--- |
|  | USA and France |  | USA and France |  |
| Rank | Trace-Test | $p$-Value | Trace-Test | $p$-Value |
| 0 | 225.130 | $[0.000]^{* *}$ | 221.980 | $[0.000]^{* *}$ |
| 1 | 158.190 | $[0.016]^{*}$ | 152.790 | $[0.036]^{*}$ |
| 2 | 111.620 | $[0.113]$ | 107.810 | $[0.178]$ |
| 3 | 75.429 | $[0.311]$ | 70.517 | $[0.490]$ |
| 4 | 47.224 | $[0.513]$ | 44.325 | $[0.679]$ |
| 5 | 27.275 | $[0.668]$ | 27.285 | $[0.667]$ |
| 6 | 11.364 | $[0.850]$ | 13.676 | $[0.688]$ |
| 7 | 3.867 | $[0.759]$ | 3.686 | $[0.783]$ |

Note: The Johansen trace test is calculated with a trend restricted to the cointegration space The number of lags in the VAR is 1 for all cases. "**" and "**" signify significance at the $1 \%$ and $5 \%$ level, respectively.

### 6.3. Estimates Using the One-Step Procedure

We now consider the one-step estimation of the model employing maximum likelihood estimation via the Kalman Filter recursions with the model specified on state space form. This method theoretically improves the efficiency as it avoids the issue from the two-step estimator of ignoring the estimation error from the first step in the second step. The estimation for US is based on the assumption of a $\operatorname{VAR}(1)$ in first differences for the transition dynamics and for France it is based on the cointegrated VAR model with two cointegrating relations. These specifications of the transient dynamics are chosen in accordance with the results reported in Section 6.2. Table 5 reports the estimated shape parameters and their standard errors. It is seen that the loading parameters and the variance are very similar to those obtained from the two-step procedure. Furthermore, the standard errors of the estimated
shape parameters are found to be very close to those of the two-step method (sometimes smaller). This indicates that small efficiency gains can be obtained by using the one-step procedure.

Table 5. Estimated loading function parameters and standard errors from the one-step procedure for French and US men and women. For US, the VAR(1) model in first difference is assumed for the transition dynamics and, for France, a VECM with two cointegrating relations is assumed. The standard errors are calculated using the inverse Fisher information criterion.

|  |  | Men |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $\lambda_{1}$ | $\lambda_{2}$ | $\lambda_{3}$ | k | $\sigma^{2}$ |
| 岂 | Estimate | 0.556 | 12.151 | 1.094 | 20.324 | 0.021 |
|  | Std. Err | 0.013 | 0.020 | 0.004 | 0.002 | 0.010 |
|  | Estimate | 0.624 | 10.809 | 1.103 | 20.014 | 0.018 |
|  | Std. Err | 0.013 | 0.020 | 0.003 | 0.002 | 0.019 |
|  |  | Women |  |  |  |  |
| 出 | Estimate | 0.648 | 18.711 | 1.453 | 19.450 | 0.024 |
|  | Std. Err | 0.013 | 0.048 | 0.003 | 0.004 | 0.018 |
| $\underset{\sim}{\infty}$ | Estimate | 0.608 | 18.930 | 1.295 | 18.703 | 0.013 |
|  | Std. Err | 0.010 | 0.043 | 0.003 | 0.004 | 0.005 |

Figure 7 shows the estimated factors (or states) for the one-step state space estimation procedure for US based on the VAR(1) specification in differences. The factor estimates for the VECM specification for France are shown in Figure 8.

When comparing the estimated factors with those obtained in the first step of the two-step approach, the results appear similar. However, the factors from the one-step estimation show a smoother development because the one-step procedure directly accounts for the transition dynamics in the estimation.


Figure 7. The factors $\kappa_{i, t}, i=0,1,2,3$ estimated by the one-step procedure for USA from 1950-2014 and assuming a $\operatorname{VAR}(1)$ model for the first difference of the factors, for both genders.


Figure 8. The factors $\kappa_{i, t}, i=0,1,2,3$ are estimated by the one-step procedure for France using data from 1950-2014 and assuming a VECM specification for the factors, for both genders.

### 6.4. Model Fit

We now compare the PFM with the LC model in terms of in-sample fit.
As the PFM does not include a constant for each age-specific death rate, we are interested in whether the model can capture the mean by relatively few parameters. As seen in Figure 9a-d, the model captures the mean well for all populations. Note that, by construction, $\alpha_{x}$ in the LC model is equal to the mean of the age specific log death rates, which corresponds to the data mean levels in the figures.

To further quantify the model fit, we calculate a pseudo $R^{2}$ for each age group by running a regression of the age-specific death rates on a constant and the fitted values. ${ }^{5}$ The pseudo $R^{2}$ 's shown in Figure 10a-d display that both the LC and the PFM fit the observed data well. However, the PFM tends to perform better around the accident hump, where the LC model is found to have poor performance.

Next, we investigate how each of the factors contribute to the explanatory power of the model by calculating the partial correlation between the log mortality and a particular factor after adjusting for the influence of the fit obtained from the remaining factors. This adjustment is necessary because the factors are non-orthogonal. Figure 11a-d display the partial correlations in excess of a $65 \%$ threshold for all ages to identify where the different factors improve the fit.

It is seen that the infant mortality factor significantly improves the fit for infants as desired. The level factor substantially improves the performance for most ages, and the accident hump factor primarily affects the mortality in the years around the accident hump. Finally, the adult factor primarily improves the fit for the adult ages as desired, but its partial explanatory power is of a smaller magnitude compared with the other factors, mainly because the adult factor is highly correlated with the factor common to all age groups.

[^37]

Figure 9. The mean of the data and the mean of the parametric factor model estimated using the two-step procedure for both men and women, for France and USA. The estimation period is 1950-2014.


Figure 10. The pseudo $R^{2}$ (within) for the PFM and the LC model for all ages estimated using the two-step procedure. The $R^{2}$ is shown for both men and women in France and the USA, respectively.


Figure 11. The partial $R^{2}$ for the infant, level, accident hump, and adult factor estimated using the two-step procedure for the years 1950-2014. The relative improvements from each of the factors are shown in excess of a $65 \%$ threshold. This is shown for both genders for France and the US, respectively.

## 7. Forecast Evaluation

In this section, we investigate the forecast performance of the PFM and compare with relevant benchmark models. For forecast evaluation and comparison, we use the Model Confidence Set (MCS) approach developed in Hansen et al. (2011). ${ }^{6}$

The MCS procedure is a test for predictive ability across a number of competing models, which sequentially removes the model that performs significantly worse than the remaining models left in the model confidence set. The procedure delineates the set of best performing models at a given confidence level among which we cannot say that any of the other models perform statistically better.

Hence, the MCS does not necessarily pick out a single best model but rather delineates a set of best models as the available information might not be able to discriminate between these models. The MCS procedure returns $p$-values, $\hat{p}_{i}$, for each model $i$ considered, and, from this, the MCS can be determined. The MCS procedure returns a $p$-value of 1 to the best performing model. ${ }^{7}$

To reduce the dimension of the forecast evaluation, we calculate the (period) life expectancy at birth which aggregates the forecasted age-specific death rates into a single measure. The (period) life

[^38]expectancy is calculated by using the standard assumption of a constant chance of death within each age interval as in Brouhns et al. (2002) ${ }^{8}$ :
\[

$$
\begin{equation*}
\bar{e}_{0}^{\uparrow}(t)=\frac{1-\exp \left(-m_{0, t}\right)}{m_{0, t}}+\sum_{k=1}^{N}\left(\prod_{j=0}^{k-1} \exp \left(-m_{j, t}\right)\right) \frac{1-\exp \left(-m_{k, t}\right)}{m_{k, t}} \tag{5}
\end{equation*}
$$

\]

where $m_{j, t}$ signifies the age-specific death rates and $\bar{e}_{0}^{\uparrow}(t)$ is the (period) life expectancy at birth.
To show the robustness of the proposed model at producing reliable forecasts, we consider data for men and women for the USA and France in the forecast evaluation. The forecasts are constructed by recursively estimating each model from 1950 onwards until year $t=1970,1971, \ldots$ and forecasting 1,10 , and 20 years ahead. This gives 43,34 , and 24 forecasts of the age-specific death rates for each model, respectively. The forecast performance is evaluated using the mean squared error of the life expectancy as the loss function. For implementation, we use the block bootstrap with a block length equal to the longest significant lag length from fitting an AR model and a confidence level of 5\%-see Hansen et al. (2011) for details.

As benchmark models, we use (1) a random walk with drift (RWD) specification for each (log) age specific death rate, (2) the Lee and Carter (1992) model with a single factor, (3) and the functional data approach (FDA) of Hyndman and Ullah (2007). Based on the analysis in Sections 5 and 6, we consider two dynamic specifications of the factor structure, a VECM (with two cointegrating relations) and a $\operatorname{VAR}(1)$ in first differences of the factors. For comparisons, we use both specifications for each country and gender estimated by the two-step procedure. For the one-step procedure, we consider estimation assuming the VECM structure for France and the VAR(1) structure in first differences for the US as found to be appropriate in Section 6. Using the two-step procedure, we further compare a $\operatorname{VAR}(1)$ model in levels and univariate ARIMA models in the dynamic specification. ${ }^{9}$ Based on the finding of unit roots and trending behaviour for each of the factors, we decide to use a random walk with drift specification as ARIMA model specification. ${ }^{10}$ For the LC model, we use a random walk with drift specification for the single factor $\kappa_{t}$. The FDA model of Hyndman and Ullah (2007) can be considered an extension of the LC model by using $K$ factors and smoothing across the death rates. ${ }^{11}$ The results are reported in Table 6 for France and in Table 7 for the USA.

[^39]Table 6. Forecasting life expectancy 1,10 , and 20 years ahead with mean-squared error criterion for US men and women evaluated using the Model Confidence Set. The mean squared error along with $p$-values for the estimated model confidence set for life expectancy. The models included in the set of best models are marked in boldface. The first five rows show the results for the parametric factor model assuming different specifications for the factor dynamics, whereas the last three rows show results for the benchmark models. The VAR1 in levels and ARIMA specifications are used for comparison.

| France |  | Men |  |  |  | Women |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | 1 Year | 10 Year | 20 Yea |  | 1 Year | 10 Year | 20 Year |
| $\sum_{a}$ |  | MSE Pval MSE Pval MSE Pval MSE Pval MSE Pval MSE Pval |  |  |  |  |  |  |
|  | VAR1 | 0.0980 .000 | 2.4930 .022 | 13.4800 | P. 000 | 0.1640 .0 | 1.7190 .000 | 7.5290 .000 |
|  | Arima | 0.1030 .001 | 0.9671 .000 | 4.1471. | 1.000 | 0.0710 .0 | 0.1440 .233 | 0.3700 .920 |
|  | $\Delta$ VAR1 | 0.0950 .006 | 0.9850 .931 | 4.6110. | . 146 | 0.0860 .0 | 0.1570 .013 | 0.4440 .415 |
|  | VECM2 | 0.0950 .002 | 0.9860 .931 | 4.5110. | 0.087 | 0.0820 .0 | 0.1380 .233 | 0.4390 .484 |
|  | VECM2SS | 0.2340 .000 | 1.3310 .284 | 4.3930. | 0.712 | 0.3590 .0 | 0.7410 .005 | 1.0410 .036 |
|  | RWD | 0.0320 .611 | 1.1350 .284 | 5.4390. | 0.000 | 0.0321 .0 | 0.1031 .000 | 0.3671 .000 |
|  | LC | 0.0990 .000 | 1.4790 .001 | 6.3670. |  | 0.1190 .0 | 0.2290 .050 | 0.4600 .329 |
|  | FDA | 0.0301 .000 | 1.0850 .875 | 5.4360. |  | 0.0370 .3 | 0.4100 .134 | 1.5810 .329 |

Table 7. Forecasting life expectancy 1,10 , and 20 years ahead with mean-squared error criterion for US men and women evaluated using the Model Confidence Set. Mean squared error along with $p$-values for the estimated model confidence set for life expectancy. The models included in the set of best models are marked in boldface.The first five rows show the results for the parametric factor model assuming different specifications for the factor dynamics, whereas the last three rows show results for the benchmark models. The VAR1 in levels and ARIMA specifications are used for comparison.

| USA |  | Men |  |  | Women |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | 1 Year | 10 Year | 20 Year | 1 Year | 10 Year | 20 Y | Year |
| $\sum_{i=1}$ |  | MSE Pval MSE Pval MSE Pval MSE Pval MSE Pval MSE Pval 0.1130 .0111 .6070 .0185 .9150 .0040 .0690 .0041 .5140 .15711 .6300 .024 |  |  |  |  |  |  |
|  | VAR1 |  |  |  |  |  |  |  |
|  | Arima | 0.1120 .013 | 0.7871 .000 | 2.4931 .000 | 0.0540 .011 | 0.5190 .225 | 1.378 | 0.017 |
|  | $\triangle$ VAR1 | 0.1100 .010 | 0.8970 .140 | 3.0940 .038 | 0.0570 .011 | 0.5620 .005 | 1.369 | 0.002 |
|  | VECM2 | 0.1270 .013 | 1.0810 .104 | 2.7650 .455 | 0.0520 .011 | 0.3291 .000 | 0.593 | 1.000 |
|  | $\triangle$ VAR1SS | 0.1170 .013 | 1.2440 .024 | 3.6620 .006 | 0.0840 .00 | 0.5750 .124 | 1.316 | 0.024 |
|  | RWD | 0.0351 .000 | 1.2400 .104 | 4.0160 .000 | 0.0231 .000 | 0.4350 .225 | 1.243 | 0.011 |
|  | LC | 0.1380 .013 | 1.7710 .018 | 5.3750 .001 | 0.0800 .006 | 0.6820 .069 | 1.790 | 0.003 |
|  | FDA | 0.0440 .154 | 1.5770 .104 | 4.8240 .006 | 0.0260 .032 | 0.4950 .225 | 1.441 | 0.024 |

France, men. For French men, the MCS using a 1-year forecast horizon includes the RWD and FDA specifications. However, when expanding the forecast horizon, the MCS now includes three variants of the PFM and, in fact, for a twenty-year forecast horizon, the MCS excludes the RWD and FDA specifications. It is interesting to observe that, in this forecast competition, the LC model is never included in the MCS. The same applies for the PFM model specification where the factors are modelled as a $\operatorname{VAR}(1)$ in levels. This is not surprising because all the factors were found to have unit roots.

USA, men. The pattern observed for French men generally applies for US men as well. However, for the 20-year horizon, only two of the PFM models are included in the MCS.

France, women. For French women and a forecasting horizon of one year, the results are rather similar to those of French men and in particular the RWD specification and the FDA model are the ones included in the MCS. For a 10-year horizon, the MCS also includes a single PFM specification and, for a 20-year horizon, only the PFM with a $\operatorname{VAR}(1)$ in levels is not included in the MCS.

USA, women. For US women, the RWD model is always in the MCS. For a 10-year horizon, the results are similar to French women and, for a 20-year horizon, the MCS is slightly smaller
than for French women and includes in particular the two PFM specifications the FDA and the RWD specifications.

In summary, the PFM class of models appears to perform especially well for longer forecast horizons and in most cases performs better than the LC model. An explanation for this result could be the structural features of the PFM class of models compared to the LC model. For longer horizons, the structural restrictions on the loadings account for different factors affecting the separate age groups. The structure implied by the PFM specification ensures a realistic shape of the mortality curve, which cannot be captured by a single factor LC model. Another conclusion is that, in situations where competing models are performing well, especially for longer horizons, the different PFM models also perform well. On the other hand, in situations where competing models are not performing so well, the PFM models are included in the MCS as seen especially for men.

## 8. Conclusions

We have suggested a multi-factor model for the term structure of mortality. The factors are identified after restrictions on the loading functions in such a way that different age groups and their factor dynamics can be addressed separately. Thus, rather than having a single factor governing all age groups as for the LC model, different factors (or trends) play a role in the way that mortality across age groups develop. In particular, we consider separate factors driving infant mortality, the accident hump mortality, mortality for the elderly in addition to a common factor affecting all age groups. We have suggested two estimation methods that are similar to estimation of term structure models considered in other contexts. In an application, we apply the methodology to mortality data for the US and France for each gender. The models are shown to provide a good fit and, for certain age groups, provides a much better fit compared to the LC model. In a forecast comparison across a range of competing models, the new class of models that we consider in the paper are shown to perform well, especially over longer forecast horizons.

Author Contributions: The authors made equal contributions.
Funding: Danish National Research Foundation: DNRF78.
Acknowledgments: The authors appreciate financial support from CREATES, Center for Research in Econometric Analysis of Time Series, funded by the Danish National Research Foundation (Grant No. DNRF78). We acknowledge helpful comments and suggestions from Siem Jan Koopman and from seminar and conference participants at CREATES lunch seminars, at Max-Planck Odense Center on the Biodemography of Aging, at the annual conference of the International Association for Applied Econometrics in Sapporo, Japan, 2017, and at Dansk Økonometrisk Selskabs annual meeting in 2016.
Conflicts of Interest: The authors declare no conflict of interest.

## References

Booth, Heather, John Maindonald, and Len Smith. 2002. Applying Lee-Carter under conditions of variable mortality decline. Population Studies 56: 325-36. [CrossRef] [PubMed]
Booth, Heather, and Leonie Tickle. 2008. Mortality modelling and forecasting: A review of methods. Annals of Actuarial Science 3: 3-43. [CrossRef]
Box, G. E., and D. R. Cox. 1964. An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological) 26: 211-52. [CrossRef]
Brouhns, Natacha, Michel Denuit, and Jeroen K. Vermunt. 2002. A poisson log-bilinear regression approach to the construction of projected lifetables. Insurance: Mathematics and Economics 31: 373-93. [CrossRef]
Byrd, Richard H., Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing 16: 1190-208. [CrossRef]
Cairns, Andrew J., David Blake, and Kevin Dowd. 2006. Pricing death: Frameworks for the valuation and securitization of mortality risk. Astin Bulletin: The Journal of the International Actuarial Association 36: 79-120. [CrossRef]

Callot, Laurent, Niels Haldrup, and Malene Kallestrup-Lamb. 2016. Deterministic and stochastic trends in the Lee-Carter mortality model. Applied Economics Letters 23: 486-93. [CrossRef]
De Jong, Piet, and Leonie Tickle. 2006. Extending Lee-Carter mortality forecasting. Mathematical Population Studies 13: 1-18. [CrossRef]
Diebold, Francis, and Robert Mariano. 1995. Comparing predictive accuracy. Journal of Business and Economic Statistics 13: 253-63.
Diebold, Francis X., and Canlin Li. 2006. Forecasting the term structure of government bond yields. Journal of Econometrics 130: 337-64. [CrossRef]
Diebold, Francis X., Glenn D. Rudebusch, and S. B. Aruoba. 2006. The macroeconomy and the yield curve: A dynamic latent factor approach. Journal of Econometrics 131: 309-38. [CrossRef]
Doornik, Jurgen A. 1998. Approximations to the asymptotic distributions of cointegration tests. Journal of Economic Surveys 12: 573-593. [CrossRef]
Doornik, Jurgen A. 1999. Erratum [approximations to the asymptotic distribution of cointegration tests]. Journal of Economic Surveys 13.
Doornik, Jurgen A. 2013. An Object-Oriented Matrix Programming Language Ox 6. Available online: https: / / ora.ox.ac.uk/objects/uuid:242f0a19-0665-4d9a-b863-774a35ce98c7 (accessed on 10 March 2019).
Doz, Catherine, Domenico Giannone, and Lucrezia Reichlin. 2011. A two-step estimator for large approximate dynamic factor models based on kalman filtering. Journal of Econometrics 164: 188-205. [CrossRef]
Durbin, James, and Siem Jan Koopman. 2012. Time Series Analysis by State Space Methods. Oxford: Oxford University Press, vol. 38.
Gavrilov, L. A., and N. S. Gavrilova. 1979. Determination of species length of life. Doklady Akademii nauk SSSR: Biological Sciences 246: 905-8.
Gavrilov, L. A., and N. S. Gavrilova. 1991. The Biology of Life Span: A Quantitative Approach. New York and Chur: Harwood Academic Publisher.
Gompertz, B. 1825. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. Philosophical Transactions of the Royal Society of London 115: 513-83. [CrossRef]
Hansen, Peter, and Asger Lunde. 2014. Mulcom 3.00. Econometric Toolkit for Multiple Comparisons.
Hansen, Peter R., Asger Lunde, and James M. Nason. 2011. The model confidence set. Econometrica 79: 453-97. [CrossRef]
Heligman, L., and J. H. Pollard. 1980. The age pattern of mortality. Journal of the Institute of Actuaries 107: 49-80. [CrossRef]
Human Mortality Database. 2016. Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available online: www.mortality.org or www.humanmortality.de (accessed on 29 June 2016).
Hyndman, Rob J. 2015. 'forecast': Forecasting Functions for Time Series And Linear Models. R package version 6.2. Available online: http:/ / github.com/robjhyndman/forecast (accessed on 13 November 2016).

Hyndman, Rob J., Heather Booth, Leonie Tickle, and John Maindonald. 2014. Demography: Forecasting Mortality, Fertility, Migration and Population Data. R package version 1.18. Available online: http:/ /CRAN.R-project. org / package=demography (accessed on 13 November 2016).
Hyndman, Rob J., and Shahid Ullah. 2007. Robust forecasting of mortality and fertility rates: A functional data approach. Computational Statistics \& Data Analysis 51: 4942-56.
Johansen, Soren. 1988. Statistical analysis of cointegration vectors. Journal of Economic Dynamics and Control 12: 231-54. [CrossRef]
Johansen, Soren. 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica: Journal of the Econometric Society 59: 1551-80. [CrossRef]
Koopman, Siem Jan, Max I. Mallee, and Michel Van der Wel. 2010. Analyzing the term structure of interest rates using the dynamic nelson-siegel model with time-varying parameters. Journal of Business \& Economic Statistics 28: 329-43.
Lazar, Dorina, and Michel M. Denuit. 2009. A multivariate time series approach to projected life tables. Applied Stochastic Models in Business and Industry 25: 806-23. [CrossRef]
Lee, Ronald D., and Lawrence R. Carter. 1992. Modeling and forecasting U.S. mortality. Journal of the American Statistical Association 87: 659-71. [CrossRef]

Lee, Ronald D., and Timothy Miller. 2001. Evaluating the performance of the Lee-Carter method for forecasting mortality. Demography 38: 537-49. [CrossRef] [PubMed]
Makeham, William Matthew. 1860. On the law of mortality and the construction of annuity tables. The Assurance Magazine, and Journal of the Institute of Actuaries 8: 301-10. [CrossRef]
McNown, Robert, and Andrei Rogers. 1989. Forecasting mortality: A parameterized time series approach. Demography 26: 645-60. [CrossRef] [PubMed]
McNown, Robert, and Andrei Rogers. 1992. Forecasting cause-specific mortality using time series methods. International Journal of Forecasting 8: 413-32. [CrossRef]
Nielsen, Bent, and Jens P. Nielsen. 2014. Identification and forecasting in mortality models. The Scientific World Journal 2014: 1-24. [CrossRef] [PubMed]
Pfaff, Bernhard. 2008. VAR, SVAR and SVEC models: Implementation within R package vars. Journal of Statistical Software 27: 1-32. Available online: http://www.jstatsoft.org/v27/i04/ (accessed on 30 January 2017). [CrossRef]
R Core Team. 2015. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. Available online: http:/ /www.R-project.org/ (accessed on 13 November 2016).
Renshaw, Arthur E., and Steven Haberman. 2003. Lee-Carter mortality forecasting with age-specific enhancement. Insurance: Mathematics and Economics 33: 255-72. [CrossRef]
Rogers, Andrei, and Jani S. Little. 1994. Parameterizing age patterns of demographic rates with the multiexponential model schedule. Mathematical Population Studies 4: 175-95. [CrossRef] [PubMed]
Siler, William. 1979. A competing-risk model for animal mortality. Ecology 60: 750-57. [CrossRef]
Stigler, Matthieu. 2010. Threshold Cointegration: Overview and Implementation in R. R package version 0.7-2. Available online: http://stat.ethz.ch/CRAN/web/packages/tsDyn/vignettes/ThCointOverview.pdf (accessed on 30 January 2017).
Strehler, Bernard L., and Albert S. Mildvan. 1960. General theory of mortality and aging. Science 132: 14-21. [CrossRef] [PubMed]
Tabeau, Ewa, Anneke van den Berg Jeths, and Christopher Heathcote. 2001. Forecasting Mortality in Developed Countries: Insights From a Statistical, Demographic and Epidemiological Perspective. Dordrecht: Kluwer Academic Publishers.
Wong-Fupuy, Carlos, and Steven Haberman. 2004. Projecting mortality trends: Recent developments in the United Kingdom and the United States. North American Actuarial Journal 8: 56-83. [CrossRef]
© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http:/ /creativecommons.org/licenses/by/4.0/).

## Article <br> <br> The Discovery of Long-Run Causal Order: <br> <br> The Discovery of Long-Run Causal Order: A Preliminary Investigation ${ }^{\dagger}$

 A Preliminary Investigation ${ }^{\dagger}$}Kevin D. Hoover ${ }^{1,2}$<br>1 Department of Economics, Duke University, Durham, NC 27708, USA; kd.hoover@duke.edu; Tel.: +1-919-660-1876<br>2 Department of Philosophy, Duke University, Durham, NC 27708, USA<br>$\dagger$ This paper arises out of a joint project with Søren Johansen, Katarina Juselius: they have been inspiring teachers, candid critics, and true friends. I am grateful for earlier collaboration with Morten Nyboe Tabor and for the skeptical, but invaluable, comments of the guest editors for the special issue, Paolo Paruolo and Rocco Mosconi, and three anonymous referees. A very early version of the paper was presented at the Econometrics Conference, Programme for Economic Modelling, University of Oxford, 1-2 September 2014. I thank the participants for valuable comments.

Received: 10 May 2018; Accepted: 23 July 2020; Published: 3 August 2020


#### Abstract

The relation between causal structure and cointegration and long-run weak exogeneity is explored using some ideas drawn from the literature on graphical causal modeling. It is assumed that the fundamental source of trending behavior is transmitted from exogenous (and typically latent) trending variables to a set of causally ordered variables that would not themselves display nonstationary behavior if the nonstationary exogenous causes were absent. The possibility of inferring the long-run causal structure among a set of time-series variables from an exhaustive examination of weak exogeneity in irreducibly cointegrated subsets of variables is explored and illustrated.


Keywords: graphical causal modeling; causal search; cointegrated vector autoregression (CVAR); weak exogeneity; irreducible cointegrating relations

JEL Classification: C32; C51; C18

> In the long run, we are all dead.
> John Maynard Keynes In the long run, we are simply in another short run.
> variously attributed Contrary to Keynes' famous dictum in the long run we are all dead, the long run is with us every day of our lives
> Walt Rostow

## 1. The Problem of Causal Order in the CVAR

Katarina Juselius and Søren Johansen's most famous contributions to econometrics, studied in detail and applied in his monograph (Johansen 1995) and in her textbook (Juselius 2006), and, jointly and singly, in a large number of journal articles, concern the cointegrated vector autoregression (CVAR). The CVAR focuses special attention on the nonstationary components and the long-run properties of the time series. The questions we address in this paper are how the long-run properties of the CVAR can be given a structural interpretation and how that interpretation might support inference of the long-run causal structure from the observable characteristics of the nonstationary data.

There are two significant traditions in time-series econometrics. ${ }^{1}$ The Cowles Commission in the 1940s and 1950s pioneered structural econometrics that conceived of the econometric problem as one of articulating and measuring economic mechanisms (Koopmans 1950; Hood and Koopmans 1953; see Morgan 1990 for a history). The articulation of mechanisms was generally referred to as the "identification problem." The major resource for securing identification was a priori economic theory. Early on, structural and causal articulation were regarded as synonymous, although subsequently causal language fell from favor (Hoover 2004). In his contribution to the 1953 Cowles volume, Simon (1953) drew on the language of experiments (actual or metaphorical) to suggest that an identified system of dynamic equations provided a map of the space of interventions in the economy. ${ }^{2}$ Simon demonstrated an isomorphism between a structurally identified model and a causally well-ordered model in a system with no stochastic variables.

A second econometric tradition, grounded more in time-series statistics, focused on process rather than structure (e.g., see Wold 1960; Granger 1969). Granger defined causality in terms of incremental predictability. Sims (1972) introduced Granger causality into empirical macroeconomics. Frequently thereafter, an equivocation between Granger's notion of causality and structural notions became commonplace. Granger himself was aware that Granger causality did not address the questions of control and counterfactual policy analysis that motivated structural understandings of causality, such as those of Simon and the Cowles Commission (Granger 1969, 1995; also White and Lu 2010, p. 194). While both the structural and the process approaches to econometrics have a concept of causation, those concepts are distinct. They may, nonetheless, be mutually informative. White and Lu (2010) and White and Pettenuzzo (2014), for instance, analyze the conditions under which Granger causality can provide information relevant to assessing structural causality (see also Hoover 2001, pp. 150-55).

The vector autoregression (VAR) arises out of the process tradition. Building on earlier criticisms of Liu (1960) and others, Sims (1980) introduced the VAR into macroeconometrics as part of a critical response to the Cowles Commission approach. Sims (1980, p. 1), attacked the structural interpretation of econometric models for using "incredible" identifying restrictions. Initially, he offered the VAR—a system of reduced-form equations in which all variables are endogenous-as a workable alternative to identified structural models.

There is a tendency to treat process accounts of causality as essentially atheoretical and data driven and to treat structural accounts as necessarily relying on a priori theory. These connections are more accidents of the history of econometrics than essential. In the case of the VAR, it rapidly became clear that reduced-form VARs were inadequate to the needs of counterfactual policy analysis-perhaps the most important use of macroeconometric models (Cooley and LeRoy 1985; Sims 1982, 1986). The structural VAR (SVAR), which imposes a causal order on the contemporaneous relationships among the endogenous variables, was seen to provide the minimum restrictions needed to identify independent shocks, which were taken to be the drivers of a dynamic system, and policy analysis was largely reduced to working out the impulse responses to those shocks (see Duarte and Hoover 2012; Hoover and Jordá 2001).

While the problem that had motivated Sims in the first place, the incredibility of the identifying restrictions, had been minimized in the SVAR, it was not eliminated; and the question, how we are to know the correct contemporaneous causal order, remains an open one. In truth, economic theory rarely provides a clear or decisive answer. In practice in most, though not all cases, SVARs were identified by assuming certain triangular causal orderings of the contemporaneous variables. Since all such causal orders are just identified, they have the same likelihood function, and, thus, there is no empirical basis for choosing among them, so long as "empirical" is restricted to likelihood information. At this point, SVAR practitioners typically claim that it is necessary to invoke prior information from economic

[^40]theory or practical institutional knowledge or common sense to pick among the equivalent causal orders. In fact, however, empirical evidence can be brought to bear on the choice. When the underlying data-generating processes (DGPs) are casually ordered in such a way that an empirically valid model of it would be over-identified, information about conditional dependence and independence among the variables in some cases will provide information that can be used to distinguish among possible causal orders. This approach has been developed with great sophistication (mainly for non-time-series data) in the so-called graphical-causal-modeling or Bayes-net literature (Spirtes et al. 2000; Pearl 2009). ${ }^{3}$ Swanson and Granger (1997) first applied a simple graphical causal search algorithm to the problem of determining the contemporaneous causal structure of an SVAR. Subsequently, more sophisticated algorithms have been applied and shown to be effective in a wide range of circumstances (Demiralp and Hoover 2003; Demiralp et al. 2008; and references therein).

Meanwhile, time-series econometrics discovered the importance of nonstationary processes and the concept of cointegration (Engle and Granger 1987). In light of these developments, the SVAR was reformulated into the CVAR. Throughout the paper, we will consider cases in which we, in fact, know the true DGP, but observe only some part of it. To be clear, our operating assumption is that a complex data-generating process governs the behavior of the economy; and the aim of structural causal modeling is to uncover a (partial) representation of the true DGP that is adequate to pragmatically required levels of detail and precision to support inter alia prediction and counterfactual analysis. ${ }^{4}$ A key question will be how much information about the DGP can be recovered from the observables.

Our interest is in long-run identification; so, we will restrict our attention to CVARs, taken to be a reduced form of a part of the economy's unobserved DGP, of the form:

$$
\begin{equation*}
\Delta \mathbf{X}_{t}=\boldsymbol{\Pi} \mathbf{X}_{t-1}+\mathbf{E}_{t}=\boldsymbol{\alpha} \boldsymbol{\beta}^{\prime} \mathbf{X}_{t-1}+\mathbf{E}_{t} \tag{1}
\end{equation*}
$$

where $\mathbf{X}=\left[x_{1}, x_{2}, \ldots x_{p}\right]^{\prime}$ is a vector of variables integrated of degree one (notated $\mathrm{I}(1)$ ), $\Pi$ is a $p \times p$ matrices of parameters; $\mathbf{E}=\left[\varepsilon_{1}, \varepsilon_{2}, \ldots \varepsilon_{p}\right]^{\prime}$ is $p$-element vector of normal residuals distributed $\mathbf{E}_{t} \sim N(\mathbf{0}$, $\Omega$ ); and $t$ subscripts indicate time. The residuals contain both unobserved causes, which we shall call "shocks," and various sorts of error. The matrix $\Omega$ is assumed to be diagonal. This assumption could be justified by economic theory or could result from orthogonalizing the residuals by multiplying through by a matrix that reflects the appropriate contemporaneous causal ordering in the manner that Choleski matrices are frequently used in the SVAR literature, a transformation that would affect the interpretation of the $\mathbf{X}_{t}{ }^{\prime}$ s.

If the variables in $\mathbf{X}$ are cointegrated (i.e., if a linear combination of nonstationary variables is itself stationary), then $\Pi$ has reduced rank ( $r$ ) and may be written as $\Pi=\boldsymbol{\alpha} \boldsymbol{\beta}^{\prime}$, where $\boldsymbol{\alpha}$ and $\beta$ are $p \times r$ matrices. Such a CVAR is said to have $r$ cointegrating relations and $q=p-r$ common trends. The rows of $\beta^{\prime}$ contain the cointegrating vectors; while the $\alpha$ matrix contains adjustment parameters. In general, the $\alpha \beta^{\prime}$ decomposition in not unique, since $\alpha$ and $\beta$ may take different values, so long as $\Pi=\alpha \beta^{\prime}$ and still remain consistent with the observations modeled in Equation (1) (Johansen 1995, p. 71; Juselius 2006, p. 216). Most of the focus in identifying the CVAR has been placed on identifying the cointegrating vectors of the $\beta^{\prime}$ on the basis of prior economic theory.

The goal of this paper is to provide a coherent account of the causal order of a CVAR and to make some preliminary suggestions about how the methods of graphical causal search in conjunction with cointegration analysis could aid in the empirical discovery of its long run, as they have already aided in the discovery of the contemporaneous causal structure.

[^41]
## 2. Graph-Theoretic Causal Order

Where other investigators have mainly focused on the cointegrating relationships encapsulated in $\beta^{\prime}$, we shift the focus to the closely related question of how trends are transmitted among the variables. Ours will be a preliminary investigation and will be restricted to cases in which all variables are $\mathrm{I}(1)$ and DGPs that can be adequately represented in a structural model that can be understood as a causally ordered consistent with a directed acyclical graph.

### 2.1. Graphs and Causal Structure

Several econometricians have given structural accounts of long-run behavior in the CVAR. They have focused mainly on the use of theory to provide the necessary identification (Davidson and Hall 1991; Pesaran and Shin 2002; Pesaran and Smith 1998; Pagan and Pesaran 2008). In contrast to economists' frequent reliance on a priori theory, in the case of stationary data, considerable headway has been made (mostly, but not entirely, outside of economics) in developing graphical causal search algorithms that can narrow the class of admissible identifications-sometimes to a unique scheme (Spirtes et al. 2000; Pearl 2009). As a preliminary to examining how some of these ideas might be extended to the nonstationary case, it will be helpful to review selectively some aspects of graphical causal analysis.

In Simon (1953) account, a structural model is a system of equations representing mechanisms in the world. ${ }^{5}$ Although the account can be generalized considerably (see Hoover 1990; 2001, chp. 3), it will do for our purposes to restrict our attention to linear equations and to treat each equation as the representation of the causal mechanism determining its left-hand-side variable (the effect) in terms of right-hand-side variables (the direct causes). The coefficients on the right-hand-side variables are taken to define the space of interventions in the causal model. Thus, an intervention, for example, to a policy rule might change the numerical value of one of the coefficients in the equation representing the rule. In a well-defined structure, the coefficients could be intervened upon independently of each other.

We analyze a restricted version of the structural approach to causality, in that it does not deal with nonlinearities, such as cross-equation restrictions, that might arise in economic optimization problems or from systemic restrictions, such as may be generated under rational expectations. In part this is a pragmatic choice to deal with the easier case first; in part, it is to maintain tighter contact with the existing graph-theoretic causal search literature; and, in part, it arises from a yet-untested conjecture that considerable empirical progress can be made with respect to long-run cause in a simple framework. The structural approach can nonetheless be further generalized; see, for example, (Hoover 1990, appendix; 2001, especially chp. 3) and White and Chalak (2009).

Graph-theoretic causal analysis represents structural systems of equations as a directed graph. The variables form the nodes or vertices of the graph, and edges connect pairs of vertices. Edges come in several forms, but we will use only one-the single-headed arrow " $\rightarrow$ ", which means "directly causes". Direct causes are also referred to as the parents of the effect or child. We restrict ourselves to directed acylical graphs (DAGs), which are adequate to the typical CVARs found in the macroeconomics literature. Graphical causal modeling is not, however, restricted to DAGs: the literature has also addressed cyclical graphs (for example, graphs in which $A$ causes $B, B$ causes $C$, and $C$ causes $A$ ) and simultaneous graphs (a particularly tight form of cyclicality in which $A$ causes $B$ and $B$ causes $A$ ) (see Richardson 1996; Phiromswad and Hoover 2013, and the references therein).

[^42]
### 2.2. Graphs and Conditional Independence

The key idea in graph-theoretic accounts of causal structure is the mapping between the causal graph and the probability distribution described of the true DGP and its reduced form. The mapping is based on Reichenbach (1956, p. 156) Principle of the Common Cause: if any two variables, $A$ and $B$, are probabilistically dependent, then either $A$ causes $B(A \rightarrow B)$ or $B$ causes $A(A \leftarrow B)$ or they have a common cause $(A \leftarrow C \rightarrow B)$. Essentially, the idea behind the principle is that correlations may not be causation, but correlations nevertheless must have a causal explanation. The Principle of the Common Cause is generalized as the "causal Markov condition" (Spirtes et al. 2000, p. 29; see also Pearl 2009, p. 30).

Without going into detail, the graph encodes certain facts of (conditional) probabilistic dependence and independence among the variables. If the data were, in fact, generated by a system of equations corresponding to the graph-as they would be, for example, in a simulation-then the joint probability distribution for those variables would embody the encoded probabilistic relations.

Some key ideas relate graphs to probabilistic independence and dependence. One variable may be a common cause of others and the effects will be rendered probabilistically independent of each other after conditioning on the common cause. Similarly, variables may stand in chains; for example, $A$ $\rightarrow C \rightarrow B$ or $A \leftarrow C \leftarrow B$. In either case, as with the common cause $(A \leftarrow C \rightarrow B)$, conditioning on the intermediating variable $C$ renders $A$ and $B$ probabilistically independent of each other. In all three cases, $C$ is said to screen (or screen-off) $A$ from $B$.

The translation of equations into graphs also generates another characteristic pattern of causal graphs. When two or more variables are causes of another variable, then several arrows will point into the effect variable. For example, $A \rightarrow C \leftarrow B$ graphs an equation in which $A$ and $B$ are the causes of $C$, and $C$ is said to be a collider on the directed path between $A$ and $B$. If $A$ and $B$, conditional on their parents, are probabilistically independent and collide at $C$, they will be probabilistically dependent conditional on $C$. With stationary data, the presence of colliders helps to orient the arrows in a graph. As we shall see presently (Section 4.2), colliders are also important to the transmission of trends, as they represent points at which new local trends are generated.

A final useful concept from graphical causality is causal sufficiency:
Definition 1. A set of variables is causally sufficient if, and only if, any variable that is excluded from the set directly causes at most one variable within the set (Spirtes et al. 2000, p. 22).

The point of invoking causal sufficiency is that the actual DGP of the economy is more complicated than any model of observable variables that an economist might analyze. When a set of variables is causally sufficient, the excluded variables are not common causes and do not induce probabilistic dependence among the observables, so that it is possible to analyze the subset of variables without loss of causal information. Clearly, causal sufficiency is a very special case that will rarely be strictly true for our models, but that sometimes might be approximately true. When it fails, we necessarily face a latent-variable problem.

Graph-theoretic search algorithms work backward from the data by systematically evaluating conditional dependence and independence relations for subsets of variables statistically and then deducing logically what graph or class of graphs or, equivalently, what econometric specifications could have generated those facts. ${ }^{6}$ We investigate the possibility of employing a strategy that was developed for stationary data to infer long-run causal structure using facts about cointegration and weak exogeneity rather than facts of causal dependence and independence.

[^43]
## 3. Where Do Stochastic Trends Come From? ${ }^{7}$

The nonstationarity of the variables in a system of equations such as Equation (1) may arise in two ways. Consider two distinct DGPs. Assume that there are two sets of variables, $\mathbf{X}_{t}$ and $\mathbf{T}_{t}$. The first corresponds to the graph Figure 1a-a simple chain:

$$
\begin{align*}
& \text { DGP 1 } \\
\Delta \mathbf{X}_{t} & =\Delta\left[\begin{array}{l}
X_{1} \\
X_{2} \\
X_{3}
\end{array}\right]_{t}=\boldsymbol{\Phi}_{\mathbf{X} \mathbf{X}} \mathbf{X}_{t-1}+\mathbf{\Phi}_{\mathbf{X T}} \mathbf{T}_{t-1}+\mathbf{E}_{t}  \tag{2}\\
& =\left[\begin{array}{ccc}
-0.2 & 0.0 & 0.0 \\
2.0 & -0.2 & 0.0 \\
0.0 & 2.0 & -0.2
\end{array}\right]\left[\begin{array}{l}
X_{1} \\
X_{2} \\
X_{3}
\end{array}\right]_{t-1}+\left[\begin{array}{l}
T \\
0 \\
0
\end{array}\right]_{t-1}+\left[\begin{array}{l}
\varepsilon_{1} \\
\varepsilon_{2} \\
\varepsilon_{3}
\end{array}\right]_{t}
\end{align*}
$$

where the $T$ s are exogenous $I(1)$ trends

$$
\begin{equation*}
\Delta \mathbf{T}_{t}=\Delta T_{t}=\mathbf{H}_{t}=\eta_{t} \tag{3}
\end{equation*}
$$

and the $\varepsilon^{\prime}$ s and the $\eta^{\prime}$ s are identically, independently distributed (i.i.d.) random shocks. The connection of DGP 1 to the CVAR of Section 2 will become clear presently.


Figure 1. (a) Causal structure of the data-generating process (DGP) 1; (b) causal structure of the DGP 2.
The second distinct DGP corresponds to the graph in Figure 1b:
DGP 2

$$
\Delta \mathbf{X}_{t}=\Delta\left[\begin{array}{l}
X_{1}  \tag{4}\\
X_{2} \\
X_{3}
\end{array}\right]_{t}=\boldsymbol{\Pi} \mathbf{X}_{t-1}+\mathbf{Z}_{t}=\left[\begin{array}{ccc}
-0.2 & 0.0 & 0.002 \\
2.0 & -0.2 & 0.000 \\
0.0 & 2.0 & -0.200
\end{array}\right]\left[\begin{array}{l}
X_{1} \\
X_{2} \\
X_{3}
\end{array}\right]_{t-1}+\left[\begin{array}{l}
\zeta_{1} \\
\zeta_{2} \\
\zeta_{3}
\end{array}\right]_{t},
$$

where the $\zeta$ s are i.i.d. random shocks.
DGP 1 shows the first of the two ways that variables may display stochastically trending behavior: $T$ trends stochastically independently of the other variables in the system because of its fundamental random-walk structure and transmits that behavior to the $X$ s, i.e., if the trend $(T)$ did not appear in Equation (2), which was otherwise unaltered (i.e., $\Phi_{X X}$ remaining the same), the system would not contain an autoregressive root of unity and the $X^{\prime}$ s be stationary).

[^44]Now suppose that the trends are latent in DGP 1, so that we observe only the Xs. To see what is implied for the cointegration of the $X s$, we can solve out $T$ to get a reduced form. The resulting system will have reduced rank (=2) and the cointegration space is spanned by two vectors given by ${ }^{8}$

$$
\left[\begin{array}{ccc}
0.0 & 2.0 & -0.2 \\
2.0 & -0.2 & 0.0
\end{array}\right] .
$$

DGP 2 shows the second way that variables can stochastically trend: here, the X's trend, not because of an exogenous cause, but because of the fine-tuning of their structural coefficients (cf. Davidson and Hall 1991, p. 239). In particular, the parameters have been chosen specifically to give DGP 2 the same cointegration properties as DGP 1. ${ }^{9}$ It is important to understand that DGP 2 is not a reduced form of DGP 1. It is a distinct structural system that happens to have coefficients that give it the same cointegration properties as DGP 1. The fact that its variables display $\mathrm{I}(1)$ trends, reflects a system property of the model that cannot be reduced to the effect of any variables that would trend without the presence of the others. In contrast, in DGP 1, the Xs trend and are cointegrated because they are driven by the same exogenous $I(1)$ trend; and that would be true whether or not the driving trend $(T)$ were observed or latent.

The I(1) behavior of the variables in DGP 2 depends on the exact values of the elements of $\Pi$. It is fragile in the sense that a small change in one of the structural coefficients that does not reflect any change in the causal graph (Figure 1b) can result in the loss of cointegration and of the trend behavior of the Xs. In contrast, DGP 1 is generic in the sense that it is robust to changes in the values of the structural coefficients (i.e., to changes that do not alter the causal graph (Figure 1a)).

To illustrate, suppose that the coefficients of DGP 2 are altered, such that the values of $\Pi$ in Equation (4) are now

$$
\Pi=\left[\begin{array}{ccc}
-0.2 & 0.0 & 0.002 \\
1.8 & -0.2 & 0.000 \\
0.0 & \mathbf{1 . 8} & -0.200
\end{array}\right]
$$

where the bold entries indicate where $\Pi$ has been altered. Now the $\operatorname{rank}(\Pi)$ is three, there is no cointegration among the variables, and, indeed, the previously nonstationary $\boldsymbol{X}_{t}$ are now stationary. ${ }^{10}$

In contrast, consider making changes of the same magnitude in the analogous part of the causal structure of DGP 1 in (2), so that

$$
\boldsymbol{\Phi}_{\mathbf{X X}}=\left[\begin{array}{ccc}
-0.2 & 0.0 & 0.0 \\
\mathbf{1 . 8} & -0.2 & 0.0 \\
0.0 & \mathbf{1 . 8} & -0.2
\end{array}\right]
$$

where again the bold numerals indicate the alterations. Unlike the case of DGP 2, qualitatively, the cointegration properties remain unchanged-there is still only the one trend, $T$, in the system. Again, if we take the trend to be latent, then, while the precise values of the cointegrating relationships have changed, the cointegration rank (2) has not. The cointegrating vectors are now

$$
\left[\begin{array}{ccc}
0.0 & 1.8 & -0.2 \\
1.8 & -0.2 & 0.0
\end{array}\right]
$$

[^45]Cointegration in DGP 2 is fragile in the sense that only specific choices of coefficients produce a trend and cointegration, and small deviations from those values can destroy those properties. Cointegration in DGP 1 is generic in the sense that small deviations in coefficients, while they alter precise values of cointegrating relations, nonetheless preserve the cointegration rank (i.e., the number of trends). The generic nature of the cointegration properties of systems like (2) is the result of the trend behavior of the $X^{\prime}$ s having an independent cause based in exogenous variables that are fundamentally I(1), while the fragility of cointegration in a DGP like (4) is the result of it arising only from the fine-tuning of the structural coefficients. Such fine-tuning could arise in specific cases for good economic reasons; however, in the spirit of Reichenbach's Principle of the Common Cause, we should assume that it would not be the general case, unless we can point to an economic explanation of why the structural coefficients take those specific values in a particular case. ${ }^{11}$ It is unlikely that cointegration generally arises from a fortuitous combination of coefficients, which combined with the fact that we often find cointegration among the observable variables without any of them being weakly exogenous, suggests that the source of nonstationary behavior and cointegration among observable variables is more typically the result of latent $\mathrm{I}(1)$ trends.

In DGP 1, we can point to specific variables that are the source of the trends. In this case, we will say that the variables are driven by genuine (or real) fundamental trends, whether those trends are themselves observed or are latent. It is conventional in the CVAR literature to say that any system of $I(1)$ variables with reduced-rank contains trends equal to the number of variables in the system less the number of cointegrating relations (the rank). These trends may generally be represented as the cumulation of the permanent shocks to the CVAR, which are backed out of the shocks to the Xs by imposing identifying assumptions (see Juselius 2006, chp. 15, especially Section 15). These representations are generally not uniquely identified, even when there are latent fundamental trends in the DGP and even when, as in DGP 2, there are no fundamental trends at all. In either case, we might call them "virtual trends," since they do not correspond to a particular variable-observable or latent.

Our working hypothesis is that trending behavior originates economically in a relatively small number of variables whose own natures are such that they are nonstationary; we call these "fundamental trends." The number of fundamental trends causally influencing a set of variables is equal to $q$ (the number of variables $(p)$ minus the rank of the $\Pi$ matrix $(r)$ ). However, the fundamental trends themselves may or may not be among the observed variables. Other variables may be nonstationary, because these fundamental trends are among their direct or indirect causes; we call these "ordinary (nonstationary) variables". In most cases, it would seem that we observe only ordinary variables, and the ultimate source of their trending behavior is to be found among their latent causes.

It might be argued that DGP 1 is also fragile because a change of parameters that rendered any of Ts stationary would upset the cointegration properties of that model in the same way that those of DGP 2 are upset by a small change in coefficients. However, that would miss the essential point. Of course, if the exogenous $T$ s were not $\mathrm{I}(1)$, then there would be no trends to transmit. The argument here, however, is that, in a structural model, it is far more likely that the source of a trend is a particular $\mathrm{I}(1)$ variable-either observed or latent-than that the source would be a group of distinct structural equations that just happen to have the right coefficients to generate what very often are multiple I(1) trends. This is ultimately not an econometric argument, but an economic one-we can more easily think of good economic reasons that a single economic variable might be a random walk (or a random walk with a drift or a random walk with a deterministic trend) than we can think of good reasons that that the parameters of several equations are appropriately tuned. For example, common sense and experience suggest that it is highly unlikely that a small change in the relative weights that a central bank places on inflation and unemployment in its reaction function would fundamentally change the

[^46]cointegration properties of a system of structural macroeconomic equations. If we do not observe such instability of the cointegration properties and we most often do not find observed exogenous I(1) variables, then it suggests that typical estimated CVARs are reduced forms and that we will have to dig deeper to discover the structure that lies behind them. Ultimately, this is an empirical hypothesis about whether CVARs based on structures like DGP 1 prove to be more economically informative than those based on structures like DGP 2. Our goal is to explore some of the implications of this hypothesis about of the typical origin of $\mathrm{I}(1)$ trends for the long-run causal structure of the world and for the possibilities of uncovering that structure (or, as least, parts of it) empirically. ${ }^{12}$

## 4. Graphical Analysis of the CVAR

The DGP that adequately represents the long-run causal structure in the economy is not directly observable. But might it be inferred on the basis of data and not simply imposed as a priori restrictions on the CVAR? We begin by showing, first, how a DGP can be represented as a causal graph; and, second, how we can think of that graph as a map of the transmission of trends through the system of variables. We then want to investigate whether the facts of cointegration and weak exogeneity among subsets of observable variables might provide the necessary empirical data to allow us to recover reliable information about the underlying DGP, analogously to the way in which graphical causal search algorithms allow us to infer the causal structure of stationary data from empirical evidence about probabilistic dependence and independence among subsets of variables. The two critical tools are Davidson (1998) analysis of irreducibly cointegrating sets of variables and Johansen (2019) state-space analysis of the CVAR, which provides an instrument for analytically determining weak exogeneity among subsets of variables. These tools allow us to explore the logic of causal inference for nonstationary data. In Section 5, we demonstrate applications of that logic that suggest a possible basis for a causal search algorithm.

### 4.1. The Canonical CVAR of a Causally Sufficient, Acyclical Graph

Consider first the long-term structure of a causally sufficient CVAR with an acyclical causal structure in which the fundamental trends are represented explicitly. In the remainder of the paper, we consider only cases for a strong form acyclicality in which we do not permit any feedback from one variable to another, even with a time delay. Thus, we rule out cases such as $X_{t} \rightarrow Y_{t+1} \rightarrow X_{t+2}$.

The DGP is given as

$$
\begin{equation*}
\Delta \xi_{t}=\boldsymbol{\Psi} \xi_{t-1}+\mathbf{H}_{t} \tag{5}
\end{equation*}
$$

where $\boldsymbol{\xi}=\left[\mathbf{X}^{\prime}, \mathbf{T}^{\prime}\right]^{\prime} ; \mathbf{T}$ is a $q \times 1$ vector of fundamental trends; $\mathbf{X}$ is a $p \times 1$ vector of ordinary variables, which may be trending (i.e., $\mathrm{I}(1))$, but are not fundamental trends; $\mathbf{H}_{t}^{\prime}=\left[\varepsilon_{1, t}, \ldots, \varepsilon_{p, t}, \eta_{1, t}, \ldots, \eta_{q, t}\right]^{\prime}$ is a $(p+q) \times 1$ vector of shocks to ordinary variables $\left(\varepsilon_{i t}, t=1,2, \ldots, p\right)$ and to fundamental trends $\left(\eta_{j t}\right.$, $j=1,2, \ldots, q)$, each of the elements of which is an identically independently distributed random shock, and $\mathbf{H}_{t} \sim \operatorname{IN}(\mathbf{0}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is diagonal.

The system can be partitioned as

$$
\Delta \xi_{t}=\left[\begin{array}{c}
\Delta \mathbf{X}  \tag{6}\\
\Delta \mathbf{T}
\end{array}\right]_{t}=\left[\begin{array}{ll}
\mathbf{\Psi}_{\mathbf{X X}} & \boldsymbol{\Psi}_{\mathbf{X T}} \\
\mathbf{\Psi}_{\mathbf{T X}} & \boldsymbol{\Psi}_{\mathbf{T T}}
\end{array}\right]\left[\begin{array}{c}
\mathbf{X} \\
\mathbf{T}
\end{array}\right]_{t-1}+\left[\begin{array}{l}
\mathbf{H}_{\mathbf{X}} \\
\mathbf{H}_{\mathbf{T}}
\end{array}\right]_{t}=\boldsymbol{\Psi} \boldsymbol{\xi}_{t-1}+\mathbf{H}_{t}
$$

where the submatrix of parameters $\boldsymbol{\Psi}_{\mathbf{X X}}$ is full rank $p \times p$, while $\boldsymbol{\Psi}_{\mathbf{X T}}$ is $p \times q, \boldsymbol{\Psi}_{\mathbf{T X}}$ is $q \times p$, and $\boldsymbol{\Psi}_{\mathbf{T T}}$ is $q$ $\times q$.

[^47]Because $\mathbf{X}$ is the vector of ordinary variables, $\boldsymbol{\Psi}_{\mathbf{X X}}$ is full rank and the eigenvalues of $\mathbf{I}_{p}+\boldsymbol{\Psi}_{\mathbf{X X}}$ must be less one in absolute value. ${ }^{13}$ If the variables in $\mathbf{T}$ are the actual $\mathrm{I}(1)$ fundamental trends, as opposed to ordinary variables that serve as the conduits of the fundamental trends into the observable system, they must be mutually causally independent, requiring $\boldsymbol{\Psi}_{\mathrm{TT}}=\mathbf{0}_{q q}$, and strongly exogenous, requiring $\boldsymbol{\Psi}_{\mathbf{T X}}=\mathbf{0}_{q p}$ (Johansen 1995, p. 77; Juselius 2006, p. 263).

The $\Psi$ matrix in (5) can be decomposed analogously to the $\Pi$ matrix in (1) such that $\Psi=\alpha \beta^{\prime}$, where $\boldsymbol{\alpha}$ is $(p+q) \times r$ and $\boldsymbol{\beta}^{\prime}$ is $r \times(p+q)$. The transitional causal structure embedded in $\boldsymbol{\Psi}$ that governs the transmission of shocks and ultimately determines the long-run causal structure reflected in (25) can be represented in this $\alpha \beta^{\prime}$-decomposition in the following canonical way-variables that are both cointegrated and directly causally connected are represented by the individual cointegrating relations expressed in $\beta$ and the effects of causes are indicated by non-zero coefficients in $\alpha$. To take a concrete example, consider a specific causal structure embedded in a DGP like (5) and represented graphically in Figure 2. (With causal time-series graphs, we suppose henceforth that the arrows correspond to a one-period lag between a direct cause and its effect.)


Figure 2. The causal structure of the DGP of cointegrated vector autoregression (CVAR) (7).
Thus, the causally canonical representation of Figure 2 would be given as

$$
\begin{align*}
& \Delta \xi_{\mathrm{t}}=\boldsymbol{\Psi} \boldsymbol{\xi}_{\mathrm{t}-1}+\mathbf{H}_{\mathrm{t}}=\left[\begin{array}{ccccccc}
\psi_{\mathrm{AA}} & 0 & 0 & 0 & 0 & \psi_{\mathrm{At}_{1}} & 0 \\
0 & \psi_{\mathrm{BB}} & 0 & 0 & 0 & \psi_{\mathrm{Bt}_{1}} & \psi_{\mathrm{Bt}_{2}} \\
0 & 0 & \psi_{\mathrm{CC}} & 0 & 0 & 0 & \psi_{\mathrm{Ct}_{2}} \\
0 & \psi_{\mathrm{DB}} & 0 & \psi_{\mathrm{DD}} & 0 & 0 & 0 \\
0 & \psi_{\mathrm{EB}} & \psi_{\mathrm{EC}} & 0 & \psi_{\mathrm{EE}} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right]\left[\begin{array}{c}
\mathrm{A} \\
\mathrm{~B} \\
\mathrm{C} \\
\mathrm{D} \\
\mathrm{E} \\
\mathrm{~T}_{1} \\
\mathrm{~T}_{2}
\end{array}\right]_{\mathrm{t}-1}+\mathbf{H}_{\mathrm{t}}  \tag{7}\\
& =\alpha \boldsymbol{\beta}^{\prime} \xi_{\mathrm{t}-1}+\mathbf{H}_{\mathrm{t}}=\left[\begin{array}{ccccc}
\alpha_{\mathrm{AA}} & 0 & 0 & 0 & 0 \\
0 & \alpha_{\mathrm{BB}} & 0 & 0 & 0 \\
0 & 0 & \alpha_{\mathrm{CC}} & 0 & 0 \\
0 & 0 & 0 & \alpha_{\mathrm{DD}} & 0 \\
0 & 0 & 0 & 0 & \alpha_{\mathrm{EE}} \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{array}\right]\left[\begin{array}{ccccccc}
1 & 0 & 0 & 0 & 0 & \beta_{\mathrm{At}_{1}} & 0 \\
0 & 1 & 0 & 0 & 0 & \beta_{\mathrm{Bt}_{1}} & \beta_{\mathrm{Bt}_{2}} \\
0 & 0 & 1 & 0 & 0 & 0 & \beta_{\mathrm{Ct}_{2}} \\
0 & \beta_{\mathrm{DB}} & 0 & 1 & 0 & 0 & 0 \\
0 & \beta_{\mathrm{EB}} & \beta_{\mathrm{EC}} & 0 & 1 & 0 & 0
\end{array}\right]\left[\begin{array}{c}
\mathrm{A} \\
\mathrm{~B} \\
\mathrm{C} \\
\mathrm{D} \\
\mathrm{E} \\
\mathrm{~T}_{1} \\
\mathrm{~T}_{2}
\end{array}\right]_{\mathrm{t}-1}
\end{align*}
$$

The rules governing the translation of the Figure 2 or any graph into the a DGP analogous to (7) are straightforward:
i. Each single-variable direct causal pair or each collider is represented by a cointegrating relationship corresponding to a unique row of the $\beta^{\prime}$ matrix where the value of the parameter for the effect is normalized to unity;

[^48]ii. There are as many adjustment parameters in $\alpha$ as there are rows in $\beta^{\prime}$ (at most one per row) with the column of each non-zero parameter in $\alpha$ corresponding to the row of one of the effects (i.e., corresponding to the row in which that variable is normalized to unity) in $\beta^{\prime}$;
iii. If any variable is a cause, but not an effect with respect to all the other variables, it corresponds to a zero row in $\alpha$ (and, thus, is weakly exogenous).

The $\beta$ matrix thus tells us which variables are related causally and, therefore, connected by edges, and the $\alpha$ matrix (equivalently the normalization of $\beta^{\prime}$ ) tells us which way the arrows point for those edges.

Except for trivial reorderings of the variables and rescalings, the DGP (7) uniquely represents the causal graph in Figure 2. Algebraically, however, the matrices $\alpha$ and $\beta$ are not unique. They can be rotated to form other pairs ( $\alpha^{*}$ and $\beta^{*}$ ) such that $\Psi=\alpha^{*} \beta^{* \prime}$. The $\alpha \beta^{\prime}$-representation and the $\alpha^{*} \beta^{* \prime}$-representation yield the same value of the likelihood function. The problem of causal search is to find empirical information, other than the value of the likelihood function, that would allow us to select the canonical representation as in DGP (7) that corresponds to the graph of the data-generating process.

### 4.2. Formation and Sharing of Local Trends

We can think of the causal graph of a system of $\mathrm{I}(1)$ variables as representing the channels of transmission of these trends. Each collider corresponds to the creation of a local trend, and the causal variables involved in the collider are cointegrated with the effect variable. The transmission of a local trend from one variable to a single other variable also implies the cointegration of the cause and the effect.

Although causal connections produce cointegration, cointegration itself is not essentially a causal notion. Instead, cointegration results either (a) when a local trend is shared by two variables or (b) whenever the number of variables sharing the same fundamental trends, whether or not they share the same local trends (i.e., whether or not they share the fundamental trends in the same proportions), exceeds the number of fundamental trends. Thus, in case (b), if there is a set of variables each of which is driven by the same $q$ fundamental trends, then any $q+1$ of them will be cointegrated. A causal connection is, thus, sufficient for the cointegration of the complete set of causes with their effect, but it is not necessary.

Proposition 1. Causal Cointegration: If each member of the set of parents of a variable $C$ in a causal graph is $I(1)$, then the set of variables consisting of $C$ and its parents, is cointegrated.

It is convenient to write the fact that a set of variables is cointegrated as $\mathrm{CI}(\mathrm{Z})$, where Z is a set of variables with two or more members. Thus, if the variables $A$ and $B$ are cointegrated, we can write this as $\mathrm{CI}(\{A, B\})$. Two terms will prove useful:

Definition 2. A cointegrating group is a set of variables in which every pair of variables shares the same common local trend-i.e., every pair is cointegrated.

Definition 3. A collider group is a set of variables consisting of a variable $C$ and the complete set of its parents.
The variables in a cointegration group share a single common local trend; while the variables in a collider group generate a new local trend at $C$. The same variable may be part of both a cointegration group and a collider group. Other sets of cointegrating variables may be in neither type of group. Davidson (1998, p. 91) introduces a useful concept, which we define here slightly differently that he does.

Definition 4. A set of variables is irreducibly cointegrating (notated IC(.)) if, and only if, it does not contain a subset that is itself cointegrated.

### 4.3. A State-Space Analysis of the CVAR

It will prove useful to examine the relationship between weak exogeneity and the causal graph. Weak exogeneity is not in itself a causal property; rather, it is a property related to the manner in which a likelihood function can be decomposed into a conditional and marginal probability distribution under a given parameterization (Engle et al. 1983). Although weak exogeneity is important because it is turns out to be the condition that guarantees that the parameters of interest can be efficiently estimated, we are not interested in the current paper in efficient estimation. Rather we want to show how zero rows in $\alpha$ in the CVAR for subsets of variables, known as "weak exogeneity" conditions, can reveal information about the causal structure of the DGP.

Given a DGP, the weak exogeneity status of its variables will depend on the model we estimate. So, for example, if (7) were the DGP with $\psi_{i j} \neq 0$ and we estimated a CVAR with precisely the form of the DGP with $\psi_{i j}$ unrestricted, then the variables $T_{1}$ and $T_{2}$ would be weakly exogenous in the model for $\left\{A, B, C, D, E, T_{1}, T_{2}\right\}_{t+1}$ given $\left\{A, B, C, D, E, T_{1}, T_{2}\right\}_{t}$ for the coefficients $\psi_{i j}$ or $\left(\alpha_{j i}, \boldsymbol{\beta}_{i j}\right), i=1,2, \ldots$, $5, j=1,2, \ldots, 7$. Our main interest, however, will be in the case in which only a subset of the variables is observed-leaving other variables in the DGP latent. So, for example, we might consider data generated by (7) but observe only B, C, and $E$. These variables can be modeled in a CVAR form, but the coefficients of the model will not in general be the same as those of (7), though we could compute them if we knew the DGP. Still, we can ask the question whether we can decompose the likelihood function of this model, with some unobserved variables, in a manner that renders some of the observed variables weakly exogenous with respect to the coefficients of a conditional model for the remaining observable variables.

We can notate this weak exogeneity using a new symbol " $\mapsto$ ", which means "is weakly exogenous for" and is to be distinguished from " $\rightarrow$," which means "directly causes." Thus, $X \mapsto Y$ can be read as "the variables in the set $X$ are weakly exogenous for the coefficients of a CVAR model of $Y$ conditional on $X$ " or, leaving the relativity to a particular set of parameters implicit, " $X$ is weakly exogenous for $Y$." If we know the causal graph of the DGP, then we can read the various weak exogeneity relationships for models of different subsets of variables from information in the causal graph. As a result, if we can identify weak exogeneity relationships for different subsets, we may be able to work backwards to determine which causal graphs could have generated them. ${ }^{14}$

The object of the analysis is to use tests of long-run weak exogeneity in CVARs of the form of Equation (1) applied to only the observable variables to discover restrictions on allowable causal ordering of the underlying DGP (6). Long-run weak exogeneity corresponds to a zero row in the $\alpha$ matrix of the CVAR, so a critical goal is, given a particular DGP, to determine what it implies for the $\alpha$ matrix of a CVAR of the subset of observable variables (Johansen 1995, Section 8.2.1; and Juselius 2006, Section 11.1).

Johansen (2019) provides a state-space analysis of the DGP of a CVAR that allows us to determine analytically what statistical tests of weak exogeneity should find (given sufficient data and so forth) for different subsets of observable variables. Fundamental trends are assumed to be latent. In order to analyze weak exogeneity among subsets of variables, Johansen partitions the ordinary variables $\mathbf{X}_{t}=\left[\mathbf{X}_{1 t}, \mathbf{X}_{2 t}\right]$ into those that are in the subset of interest $\mathbf{X}_{1 t}$ (referred to as observed) and those outside the subset $\boldsymbol{X}_{2 t}$ (referred to as the unobserved). Then, rather than partitioning $\boldsymbol{\Psi}$ as in (6), partition

[^49]it as $\boldsymbol{\Psi}=\left[\begin{array}{cc}p \times p & p \times m \\ \mathbf{M} & \mathbf{C} \\ 0 & 0 \\ m \times p & m \times m\end{array}\right]$, where the submatrices of parameters may or may not coincide with the $\boldsymbol{\Psi}_{\mathrm{ij}}$, depending on whether any ordinary variables are unobserved. The $m \times p$ null element in the lower left-hand corner of the $\Psi$ matrix corresponds to the assumption that the fundamental trends are strongly exogenous, and the $m \times m$ null element in the lower right-hand corner indicate that fundamental trends do not cause one another.

$$
\text { The submatrix } \underset{p \times p}{\mathbf{M}}=\left[\begin{array}{lll}
p_{1} \times p_{1} & p_{1} \times p_{2} \\
\mathbf{M}_{11} & \mathbf{M}_{12} \\
\mathbf{M}_{21} & \mathbf{M}_{22} \\
p_{2} \times p_{1} & p_{2} \times p_{2}
\end{array}\right] \text { contains the parameters of the ordinary variables. Only }
$$ the parameters in $\mathbf{M}_{11}$ relate exclusively to the $p_{1}$ observed ordinary variables, while the other $\mathbf{M}_{i j}$ contain parameters that relate partly or exclusively to the $p_{2}$ latent ordinary variables. The submatrix $\underset{p \times m}{\mathbf{C}}=\left[\begin{array}{cc}\mathbf{C}_{1 \times p_{1}}^{\prime} & \mathbf{C}^{\prime}{ }_{2}^{2} \\ m \times p_{2}\end{array}\right]^{\prime}$ contains the coefficients in $\mathbf{C}_{1}$ that relate to the effects of the latent fundamental trends on the observed ordinary variables and those in $\mathrm{C}_{2}$ that relate to the their effects on the unobserved ordinary variables.

A state-space representation of DGP (6) can then be given.

$$
\begin{gather*}
\Delta \mathbf{X}_{1 t+1}=\mathbf{M}_{11} \mathbf{X}_{1 t}+\mathbf{M}_{12} \mathbf{X}_{2 t}+\mathbf{C}_{1} \mathbf{T}_{t}+\boldsymbol{\varepsilon}_{1 t+1} ;  \tag{8}\\
\Delta \mathbf{X}_{2 t+1}=\mathbf{M}_{21} \mathbf{X}_{1 t}+\mathbf{M}_{22} \mathbf{X}_{2 t}+\mathbf{C}_{2} \mathbf{T}_{t}+\varepsilon_{2 t+1} ;  \tag{9}\\
\Delta \mathbf{T}_{t+1}=\boldsymbol{\eta}_{t+1} \tag{10}
\end{gather*}
$$

where $t=0,1, \ldots, n-1$, and $\mathbf{T}_{0}=\mathbf{0}$ and $\mathbf{X}_{0}=\mathbf{0}$. The shocks are partitioned into those affecting ordinary variables $(\varepsilon)$ and those affecting the latent variables $(\eta)$, with $\left(\varepsilon_{t}, \eta_{t}\right) \sim$ i.i.d. $N_{p+m}(0, \Omega)$, $\Omega=\left[\begin{array}{ccc}\Omega_{\varepsilon 1} & 0 & 0 \\ 0 & \Omega_{\varepsilon 2} & 0 \\ 0 & 0 & \Omega_{\eta}\end{array}\right]$, where $\Omega$ is diagonal. In keeping with the distinction between ordinary variables and fundamental trends, we assume that the eigenvalues of $\mathbf{I}_{p}+\mathbf{M}, \mathbf{I}_{p 1}+\mathbf{M}_{11}$, and $\mathbf{I}_{p 1}+\mathbf{M}_{22}$ are less than one in absolute value, so that the source of the nonstationarity of $\mathbf{X}_{t}$ is the fundamental trends rather than its own dynamics.

The matrix $\mathbf{C}$ represents the proportions of fundamental trends present in observable variables but transmitted to them through latent causal connections and not via causal relationships among the observable variables. Thus, while the non-zero entries of $\mathbf{M}$ correspond to the edges in a causal graph, C is not given a direct graphical interpretation. The fundamental trends are embedded in $\mathbf{T}$, but the variables included in $\mathbf{T}$ should be regarded as local trends, which may either be latent fundamental trends directly causing the observed variables or latent ordinary variables that carry some linear combination of fundamental trends and cause the observable variables. Therefore, while we have assumed that $\Omega_{\eta}$ is diagonal, it need not be (and the conclusions about weak exogeneity in the next subsection would be unaffected).

Suppose that the DGP is described as in systems (8)-(10), and we wish to know whether any of the observed variables $\left(\mathbf{X}_{1 t}\right)$ are weakly exogenous in a CVAR of the observed variables only. This comes down to the question of whether $\alpha$ in that CVAR has any zero rows. Johansen proves that the $\alpha$ of such a CVAR can be written as

$$
\begin{equation*}
\boldsymbol{\alpha}=\boldsymbol{\Sigma}\left(\mathbf{M}_{12} \mathbf{V}_{2 \mathbf{T}}+\mathbf{C}_{1} \mathbf{V}_{\mathbf{T T}}\right)_{\perp}, \tag{11}
\end{equation*}
$$

where the conditional variances are

$$
\mathbf{V}=\operatorname{var}\left[\left.\left[\begin{array}{c}
\mathbf{X}_{2 t} \\
\mathbf{T}_{t}
\end{array}\right] \right\rvert\, \mathbf{X}_{1 t}\right]=\left[\begin{array}{ll}
\mathbf{V}_{22} & \mathbf{V}_{2 \mathbf{T}} \\
\mathbf{V}_{\mathbf{T} 2} & \mathbf{V}_{\mathbf{T T}}
\end{array}\right]
$$

and the long-run variances are

$$
\boldsymbol{\Sigma}=\operatorname{var}\left(\mathbf{X}_{1 t}\right)=\left[\begin{array}{ll}
\mathbf{M}_{12} & \mathbf{C}_{1}
\end{array}\right] \mathbf{V}\left[\begin{array}{c}
\mathbf{M}_{12}^{\prime} \\
\mathbf{C}_{1}^{\prime}
\end{array}\right]+\Omega_{\varepsilon \in 1} ;
$$

see (Johansen 2019, Sections 2 and 3, especially Equations (12) and (13), Theorem 3, and Equation (18)).
In the simpler case, in which all variables are observed (i.e., there are no $X_{2}$ 's), Johansen (2019, Section 3, Case 1) shows the formula in Equation (11) can be made even simpler:

$$
\begin{equation*}
\alpha=\Omega_{\varepsilon 1} \mathrm{C}_{1 \perp} . \tag{12}
\end{equation*}
$$

### 4.4. Weak Exogeneity and Causal Order

Johansen (2019) state-space representation and his Theorem 2 offer a tool for analyzing weak exogeneity for subsets of variables in the DGP. These, in turn, correspond in systematic ways to facts about the causal structure of the DGP itself. Consider some illustrative cases:

Case 1. Consider the causal graph in Figure 3, in which all ordinary variables are observed and only the fundamental trends are unobserved, so that (12), the simpler formula for $\alpha$, applies. The DGP in Equations (8)-(10) specializes to

$$
\begin{align*}
& \Delta \mathbf{X}_{1 t+1}=\Delta\left[\begin{array}{l}
A \\
B \\
C
\end{array}\right]_{t+1} \\
&=\left[\begin{array}{ccc}
\psi_{A A} & 0 & 0 \\
0 & \psi_{\mathrm{BB}} & 0 \\
\psi_{\mathrm{CA}} & \psi_{\mathrm{CB}} & \psi_{\mathrm{CC}}
\end{array}\right]\left[\begin{array}{l}
A \\
B \\
C
\end{array}\right]_{t}+\left[\begin{array}{ll}
\psi_{A T_{1}} & \psi_{A T_{2}} \\
\psi_{B T_{1}} & \psi_{B T_{2}} \\
0 & 0
\end{array}\right]\left[\begin{array}{l}
T_{1} \\
T_{2}
\end{array}\right]_{t}+\left[\begin{array}{l}
\varepsilon_{A} \\
\varepsilon_{B} \\
\varepsilon_{C}
\end{array}\right]_{t+1}=\mathbf{M}_{11} \mathbf{X}_{1 t}+\mathbf{C}_{1} \mathbf{T}_{t}+\varepsilon_{1 t+1}  \tag{13}\\
& \Delta \mathbf{T}_{t+1}=\Delta\left[\begin{array}{l}
T_{1} \\
T_{2}
\end{array}\right]_{t+1}=\boldsymbol{\eta}_{t+1}=\left[\begin{array}{l}
\eta_{1} \\
\eta_{2}
\end{array}\right]_{t+1}, \tag{14}
\end{align*}
$$

where

$$
\Omega_{\varepsilon 1}=\left[\begin{array}{ccc}
\omega_{A A} & 0 & 0 \\
0 & \omega_{B B} & 0 \\
0 & 0 & \omega_{C C}
\end{array}\right]
$$

where $\omega_{\mathrm{ii}}=\operatorname{var}\left(\varepsilon_{\mathrm{it}}\right), \mathrm{i}=\mathrm{A}, \mathrm{B}, \mathrm{C}$; Thus,

$$
\boldsymbol{\alpha}=\Omega_{\varepsilon \in 1} \mathbf{C}_{1 \perp}=\left[\begin{array}{ccc}
\omega_{A A} & 0 & 0 \\
0 & \omega_{B B} & 0 \\
0 & 0 & \omega_{C C}
\end{array}\right]\left[\begin{array}{cc}
\psi_{14} & \psi_{15} \\
\psi_{24} & \psi_{25} \\
0 & 0
\end{array}\right]_{\perp}=\left[\begin{array}{ccc}
\omega_{A A} & 0 & 0 \\
0 & \omega_{B B} & 0 \\
0 & 0 & \omega_{C C}
\end{array}\right]\left[\begin{array}{l}
0 \\
0 \\
1
\end{array}\right]=\left[\begin{array}{l}
0 \\
0 \\
*
\end{array}\right],
$$

where the asterisk ( ${ }^{*}$ ) indicates a non-zero value. ${ }^{15}$ The first two rows of $\boldsymbol{\alpha}$ are zero and, therefore, $A$ and $B$ are weakly exogenous for $C$ (i.e., $\{A, B\} \mapsto C$ ). Notice that it does not matter, what the causal relations are among the observables, since they are encoded in the $\mathbf{M}_{11}$ matrix, which plays no part in the determination of $\boldsymbol{\alpha}$ in Equation (11). What matters is which variables convey the fundamental trends to the observables.
Case 2. Unfortunately, the simple mapping between weak exogeneity and causal connection suggested by Case 1 does not hold up. Consider Figure 4, which adds the variable D and edges connecting it to other variables in Figure 3. The analysis proceeds just as in Case 1. Again, since all variables

[^50]are observable, the simpler formula (12) applies. The other relevant matrices of the state-space formulation are given by
\[

\mathbf{X}_{t}=\left[$$
\begin{array}{c}
A \\
B \\
C \\
D
\end{array}
$$\right]_{t}, \mathbf{C}_{1}=\left[$$
\begin{array}{cc}
\psi_{15} & \psi_{16} \\
\psi_{25} & \psi_{26} \\
0 & 0 \\
0 & \psi_{46}
\end{array}
$$\right] , and \Omega_{\varepsilon} 1=\left[$$
\begin{array}{cccc}
\omega_{A A} & 0 & 0 & 0 \\
0 & \omega_{B B} & 0 & 0 \\
0 & 0 & \omega_{C C} & 0 \\
0 & 0 & 0 & \omega_{D D}
\end{array}
$$\right] .
\]

These imply that

$$
\alpha=\Omega_{\varepsilon 1} C_{1 \perp}=\left[\begin{array}{cc}
0 & * \\
0 & * \\
* & 0 \\
0 & *
\end{array}\right]
$$

which has no zero rows; which, in turn, implies that none of the variables is weakly exogenous. ${ }^{16}$ The variables $A, B, C, D$ are cointegrated $(\mathrm{CI}(\{A, B, C, D\}))$; but with two fundamental trends and four variables, every three-member subset of the ordinary variables is also cointegrated, implying not IC $(\{A, B, C, D\})$. This appears to be a robust finding-the parents in a collider are weakly exogenous only when the colliding set is irreducibly cointegrated.
Case 3. It is tempting to think that we might consider an irreducible subset of the variables in Figure 4, such as $\{A, B, C\}$ and find the same weak exogeneity relations as we did in Figure 3. That, however, does not work. In analyzing the subset, we are effectively treating $D$ as an unobserved variable; and we must, therefore, apply the more general formula (11), which requires additional information. The critical elements of the state-space representation of this reduced system are

$$
\begin{gathered}
\mathbf{X}_{1 t}=\left[\begin{array}{c}
A \\
B \\
C
\end{array}\right]_{t} ; \mathbf{x}_{2 t}=[D]_{t} ; \\
\mathbf{M}_{12}=\left[\begin{array}{c}
0 \\
0 \\
\psi_{34}
\end{array}\right] ; \mathbf{C}_{1}=\left[\begin{array}{cc}
\psi_{A T_{1}} & \psi_{A T_{2}} \\
\psi_{B T_{1}} & \psi_{B T_{2}} \\
0 & 0
\end{array}\right] ; \Omega_{\varepsilon}=\left[\begin{array}{ccc}
\omega_{A A} & 0 & 0 \\
0 & \omega_{A A} & 0 \\
0 & 0 & \omega_{A A}
\end{array}\right] ;
\end{gathered}
$$

and

$$
\mathbf{V}=\operatorname{var}\left[\left.\begin{array}{c|}
\mathbf{X}_{2 t} \\
\mathbf{T}_{t}
\end{array} \right\rvert\, \mathbf{X}_{1 t}\right]=\left[\begin{array}{ll}
\mathbf{V}_{22} & \mathbf{V}_{2 \mathbf{T}} \\
\mathbf{V}_{\mathbf{T} 2} & \mathbf{V}_{\mathbf{T T}}
\end{array}\right]=\left[\begin{array}{ccc}
* & * & * \\
* & * & * \\
* & * & *
\end{array}\right] .
$$

(Note that, although $\Omega_{\eta}$ is diagonal by assumption, the off-diagonal elements of $\mathbf{V}_{\mathbf{T T}}$ here are nonzero. This is the result of $D$, transmitting $T_{2}$ to the collider at $C$. The calculation of $\mathbf{V}$ (see Equation (11)) conditions $\left\{D, T_{1}, T_{2}\right\}$ on $\{A, B, C\}$ and, in effect, conditions the independent (distal)

[^51]causes $T_{1}$ and $T_{2}$ on their common (indirect) effect, which induces probabilistic dependence between them.) The variance of the $\mathbf{X}_{1 t}$ is
\[

$$
\begin{aligned}
\boldsymbol{\Sigma}=\operatorname{var}\left(\mathbf{X}_{1 t}\right) & =\left[\begin{array}{cc}
\mathbf{M}_{12} & \mathbf{C}_{1}
\end{array}\right] \mathbf{V}\left[\begin{array}{c}
\mathbf{M}_{12}^{\prime} \\
\mathbf{C}_{1}^{\prime}
\end{array}\right]+\mathbf{\Omega}_{\boldsymbol{\varepsilon} 1} \\
& =\left[\begin{array}{ccc}
0 & \psi_{A T_{1}} & \psi_{A T_{2}} \\
0 & \psi_{B T_{1}} & \psi_{B T_{2}} \\
\psi_{C D} & 0 & 0
\end{array}\right]\left[\begin{array}{ccc}
* & * & * \\
* & * & * \\
* & * & *
\end{array}\right]\left[\begin{array}{ccc}
0 & 0 & \psi_{C D} \\
\psi_{A T_{1}} & \psi_{B T_{1}} & 0 \\
\psi_{A T_{2}} & \psi_{B T_{2}} & 0
\end{array}\right]+\left[\begin{array}{ccc}
\omega_{A A} & 0 & 0 \\
0 & \omega_{B B} & 0 \\
0 & 0 & \omega_{C C}
\end{array}\right] \\
& =\left[\begin{array}{ccc}
* & * & * \\
* & * & * \\
* & * & *
\end{array}\right]
\end{aligned}
$$
\]

and

$$
\begin{aligned}
\alpha=\Sigma\left(\mathbf{M}_{12} \mathbf{V}_{2 \mathbf{T}}+\mathbf{C}_{1} \mathbf{V}_{\mathbf{T T}}\right)_{\perp} & =\Sigma\left(\left[\begin{array}{l}
0 \\
0 \\
*
\end{array}\right]\left[\begin{array}{ll}
* & *
\end{array}\right]+\left[\begin{array}{ll}
* & * \\
* & * \\
0 & 0
\end{array}\right]\left[\begin{array}{ll}
* & * \\
* & *
\end{array}\right]\right)_{\perp} \\
& =\Sigma\left[\begin{array}{ll}
* & * \\
* & * \\
0 & *
\end{array}\right]_{\perp}=\left[\begin{array}{lll}
* & * & * \\
* & * & * \\
* & * & *
\end{array}\right]\left[\begin{array}{c}
* \\
* \\
*
\end{array}\right]=\left[\begin{array}{l}
* \\
* \\
*
\end{array}\right] .
\end{aligned}
$$

With no zero rows in $\alpha$, none of the variables is weakly exogenous. Although $D$ is unobservable in the DGP that actually determines the value of the observable variables, it provides a conduit from the fundamental trends to $C$ that is distinct from the observable conduits, $A$ and $B$. It is as if the graph of Figure 4 has been transformed into Figure 6, where the dashed arrow indicates a causal connection between $T_{2}$ and $C$, mediated by $D$ in the DGP but not observable in the CVAR of the subset $\{A, B, C\}$. Unobserved mediating causes, like D , can make an indirect causal connection appear to be direct.
Case 4. In Case 3, weak exogeneity failed to obtain, even though the causal connections were genuine. It can also happen that weak exogeneity does obtain, even when causal connections are missing. Consider Figure 5. The graph shows not $(A \rightarrow C)$ and $\operatorname{not}(B \rightarrow D)$ and $n o t(B \rightarrow E)$, although $B$ does indirectly cause $E$. Using the same state-space methods, but omitting the details here, we can show that $\{A, B\} \mapsto\{C, D, E\}$. And, looking at subsets of variables $\{A, B\} \mapsto D$. Thus, $\{A, B, D\}$ have the same apparent pattern of weak exogeneity as found for $\{A, B, C\}$ in Case 1 (Figure 3); yet these variables do not form a collider group in Figure 5. But notice $\mathrm{CI}(\{A, B, D\})$, but also $\mathrm{CI}(\{A, D\})$. The set $\{A, B, \mathrm{D}\}$, therefore, is not irreducibly cointegrated. It appears that a mapping between weak exogeneity and causal connections can be established only in irreducibly cointegrated sets.
Case 5. Weak exogeneity may fail to track direct cause. Consider a causal chain:

$$
T \rightarrow A \rightarrow B \rightarrow C \rightarrow D
$$

All four observable variables form a single cointegration group, sharing the single fundamental trend. Note that $B \mapsto C$ and that $\{B, C\}$ form a cointegration group. We might be tempted to conclude that these facts would warrant inferring what is, in fact, true that $B \rightarrow C$. A similar case shows the problem: $A \mapsto C$ and $\mathrm{CI}(\{A, C\})$; but, in fact, it is not true that $A \rightarrow \mathrm{C}(A$ is an indirect, but not a direct, cause of $C$ ). It is worth showing why it is the case that $A \mapsto C$, as it highlights a subtle issue. We take $\{A, C\}$ to be observed and $\{B, D\}$ to be unobserved. Then the relevant matrices are

$$
\begin{gathered}
\mathbf{X}_{1 t}=\left[\begin{array}{l}
A \\
C
\end{array}\right]_{t} ; \mathbf{X}_{2 t}=\left[\begin{array}{c}
B \\
D
\end{array}\right]_{t} \\
\mathbf{M}_{12}=\left[\begin{array}{cc}
0 & 0 \\
\psi_{B C} & 0
\end{array}\right] ; \mathbf{C}_{1}=\left[\begin{array}{c}
\psi_{A T} \\
0
\end{array}\right] ;
\end{gathered}
$$

$$
\mathbf{V}=\operatorname{var}\left[\left.\begin{array}{c}
\mathbf{X}_{2 t} \\
\mathbf{T}_{t}
\end{array} \right\rvert\, \mathbf{X}_{1 t}\right]=\left[\begin{array}{ll}
\mathbf{V}_{22} & \mathbf{V}_{2 \mathbf{T}} \\
\mathbf{V}_{\mathbf{T} 2} & \mathbf{V}_{\mathbf{T T}}
\end{array}\right]=\left[\begin{array}{ccc}
* & 0 & 0 \\
0 & * & 0 \\
0 & 0 & *
\end{array}\right] ; \mathbf{\Omega}_{\boldsymbol{\varepsilon} 1}=\left[\begin{array}{cc}
\omega_{A A} & 0 \\
0 & \omega_{\mathrm{CC}}
\end{array}\right]
$$



Figure 3. Causal structure of the DGP of Case 1.


Figure 4. Causal structure of the DGP of Case 2.


Figure 5. Causal structure of the DGP of Case 4.


Figure 6. Virtual causal structure of Case 2 when $D$ is unobserved.
The variance of the $\mathbf{X}_{1 t}$ is

$$
\begin{aligned}
\boldsymbol{\Sigma}=\operatorname{var}\left(\mathbf{X}_{1 t}\right) & =\left[\begin{array}{ll}
\mathbf{M}_{12} & \mathbf{C}_{1}
\end{array}\right] \mathbf{V}\left[\begin{array}{c}
\mathbf{M}_{12}^{\prime} \\
\mathbf{C}_{1}^{\prime}
\end{array}\right]+\mathbf{\Omega}_{\boldsymbol{\varepsilon} 1} \\
& =\left[\begin{array}{ccc}
0 & 0 & \psi_{A T} \\
\psi_{B C} & 0 & 0
\end{array}\right]\left[\begin{array}{ccc}
* & 0 & 0 \\
0 & * & 0 \\
0 & 0 & *
\end{array}\right]\left[\begin{array}{cc}
0 & \psi_{B C} \\
0 & 0 \\
\psi_{A T} & 0
\end{array}\right]+\left[\begin{array}{cc}
\omega_{A A} & 0 \\
0 & \omega_{C C}
\end{array}\right]=\left[\begin{array}{cc}
* & 0 \\
0 & *
\end{array}\right] ;
\end{aligned}
$$

and
$\boldsymbol{\alpha}=\boldsymbol{\Sigma}\left(\mathbf{M}_{12} \mathbf{V}_{2 \mathbf{T}}+\mathbf{C}_{1} \mathbf{V}_{\mathrm{TT}}\right)_{\perp}=\Sigma\left(\left[\begin{array}{ll}0 & 0 \\ * & 0\end{array}\right]\left[\begin{array}{l}0 \\ 0\end{array}\right]+\left[\begin{array}{l}* \\ 0\end{array}\right][*]\right)_{\perp}=\Sigma\left[\begin{array}{l}* \\ 0\end{array}\right]_{\perp}=\left[\begin{array}{ll}* & 0 \\ 0 & *\end{array}\right]\left[\begin{array}{l}0 \\ *\end{array}\right]=\left[\begin{array}{l}0 \\ *\end{array}\right]$.
The zero row in $\alpha$ implies that $A \mapsto C$. The result hinges crucially on $\mathbf{V}_{2 T}$ being a zero matrix. This is, in turn, implied by the fact that $A$ screens off $B$ and $D$ from $T$ in the graph. Conditioning on the screening variable $A$ as is done in the calculation of $\mathbf{V}_{2 \mathbf{T}}$ renders both $B$ and $D$ probabilistically independent of $T$.

Using a similar analysis, it is also easy to show that the subset $\{B, D\}$ displays the same pattern as $\{A, C\}: B \mapsto D$ and $C I(\{B, D\})$, yet it is not true that $B \rightarrow \mathrm{D}$. The example shows that we have to be careful in making such inferences, but not that they are hopeless. Note that we can show that $A \mapsto\{B$, $C, D\} ; B \mapsto\{C, D\}$; and $C \mapsto D$; so that the variables form a nested hierarchy with $A$ at the top. This hierarchy can be reinterpreted as a chain: $A \mapsto B$ and all variables lower in the hierarchy; $B \mapsto C$ and all variables lower in hierarchy; $C \mapsto D$; and $D$ is not weakly exogenous for any variable. Such as chain recapitulates the causal graph. The lesson is that a when a variable is weakly exogenous for another variable in a cointegration group, it is a direct cause only if it is adjacent in the sense of sitting at the immediately higher step of the hierarchy.

Although we have not provided a proof, these cases suggest how to read weak exogeneity off a causal graph. There are four conjectured criteria:
A. Within a set of variables that form a cointegration group, a particular variable is weakly exogenous for the group if, and only if, it is the sole source of the local trend that cointegrates the group;
B. The parents in any set of variables that form a collider group in which two or more local trends are combined are weakly exogenous for the child in the collider group, provided that the number of variables in the group is fewer than one plus the number of fundamental trends carried by those variables;
C. If a collider fulfills criterion B, then in any set that replaces one or more weakly exogenous parents with a variable in the same cointegration group as that parent, provided the variable is itself weakly exogenous for the parent, will also be weakly exogenous for the child. (Thus, in Figure 5, in the collider $\{A, C, E\},\{A, C\} \mapsto E$; but in the set in which $B$ replaces $C$ (both in the same collider group), $\{A, B\} \mapsto E)$ );
D. If a collider fulfills criterion $B$, then any variable that is weakly exogenous for the child, either as a parent or as a member of the same cointegration group that replaces the parent, will be weakly exogenous for a variable that replaces the child from a cointegration group that includes the child and for which it is weakly exogenous. (Thus, in Figure $2,\left\{T_{1}, T_{2}\right\} \mapsto B$, but in the set that replaces $B$ with $D$, which are both in the same cointegration group, $\left\{T_{1}, T_{2}\right\} \mapsto D$.)

The inferential lessons of Cases 1-5 can be summarized in three conjectured rules, consistent with visual reading of the graph:

Rule 1. If $A \mapsto B$, then not $B \rightarrow A$.
Rule 1 simply says that causation cannot run against the direction of weak exogeneity.
Rule 2. In a cointegration group, if $A \mapsto B$ and there is no $C$ such that $A \mapsto C$ and $C \mapsto B$, then $A \rightarrow B$.
Rule 2 says that bivariate weak exogeneity coincides with direct causation, provided that the variables are adjacent. ${ }^{17}$

Rule 3. A set of variables $\boldsymbol{W}$ with $k \leq q$ members forms a collider at one of its members (call it variable $C$ ), if (i) $I C(W)$; (ii) $W_{-C} \mapsto C$, where $W_{-C}$ is the set $W$ omitting $C$; (iii) it is not the case that any member $B \in W_{-C}$ is a member of a cointegration group $Z$ such that, for any member $D \in Z$ (excluding $B$ ), $B \mapsto D$ and $W_{-B+D} \mapsto C$, where $W_{-B+D}$ is $W$ with $D$ taking the place of $B$; and iv) it is not the case that $C$ is a member of a cointegration group $Z$ such that for any member $E \in Z$ (excluding $C$ ) that $E \mapsto C$.

Rules 3 says that if a set of $k+1$ variables is irreducibly cointegrated and $k$ variables are jointly weakly exogenous for the $k+1$ th variable, then they form a collider, provided that each of the weakly exogenous variables is adjacent to the third variable (established by conditions (iii) and (iv)).

## 5. The Basis for a Long-Run Causal Search Algorithm?

The DGP that adequately represents the causal structure in the economy is not directly observable. Might it be inferred on the basis of data and not simply imposed as a priori restrictions on the CVAR? Based on our analysis of long-run causal structure, can we recover reliable information about the underlying DGP from the facts of cointegration and weak exogeneity analogously to the way in which graphical causal search algorithms infer causal structure for stationary data from empirical evidence about probabilistic dependence and independence among subsets of variables?

Davidson (1998, Section 3) proposes a search algorithm that identifies every irreducible cointegrating set of variables within a CVAR. He then uses that information where possible to identify the cointegrating relations in the $\beta^{\prime}$ matrix. This strategy is successful in some cases and not others. There is an analogy with causal search for stationary variables. Despite the slogan, "correlation is not causation," it is sometimes possible to infer causal direction from tests of unconditional dependence. For example, for a causally sufficient set of three stationary variables with an acyclical data-generating process, if $A$ and $C$ are not correlated, but $A$ and $B$ and $B$ and $C$ are correlated, then $A \rightarrow B \leftarrow C$ is the only consistent causal graph. In most cases, however, unconditional independence is not enough. Relations of conditional dependence and independence provides a richer source of information for inferring the direction, as well as the existence of causal edges (see Section 2.2 above).

Davidson's schema places cointegration in something like the logical role of unconditional independence (or correlation) in the stationary case. The analysis of Section 4 suggests that Davidson's inferential scheme can be further developed by explicitly recognizing, first, that the ultimate source of nonstationarity in any set of variables is often found in latent trends and, second, that assessment

[^52]of weak exogeneity may provide evidence of causal asymmetry. Within irreducibly cointegrated subsets of the variables, weak exogeneity can function in something like the logical role of conditional independence, when processed according to the three rules of Section 4.4, and may provide richer, empirically grounded information about the identification of the CVAR. As with causal search in the stationary case, the application of these rules is unlikely to identify every possible causal graph but may sometimes be able to partially or completely uncover the underlying causal structure.

To illustrate, we analyze two cases-one with and one without causal sufficiency.

### 5.1. Long-Run Causal Search in a Causally Sufficient Graph

Consider the DGP in Figure 2 and assume that its variables are causally sufficient and all (including the fundamental trends) are observed. We are interested in the logic of causal inference rather than the statistical problem of inference, so we also assume that prior statistical testing has successfully identified the facts with respect to the cointegration rank of the system and cointegration and weak exogeneity among any subset of variables. (In the language of the causal search literature, we assume that we have an oracle.) Naturally, in practice our inference cannot be more certain than the statistical inferences that provide our assumed facts. Can we use this information to recover the graph of the DGP?

The inference problem can be viewed as how to place the zero and non-zero coefficients in the $\alpha$ and $\beta^{\prime}$ matrices in Equation (7).

Given that we know that the cointegration rank is 5 , we know that there are two fundamental trends. This implies that $\alpha$ is $7 \times 5$ and $\beta^{\prime} 5 \times 7$. Since $T_{1}$ and $T_{2}$ are weakly exogenous with respect to all other variables in the system, we may conclude that, even if they are not identical with the fundamental trends (which in this case, of course, they are), they are at least the unique sources introducing those trends into the system. And we are entitled to enter zeroes in the entire rows of $\alpha$ corresponding to $T_{1}$ and $T_{2}$. Without loss of generality, we may enter non-zero $\boldsymbol{\alpha}_{i j} \mathrm{~s}$ along the main diagonal of the submatrix of $\alpha$, excluding the $T_{1}$ - and $T_{2}$-rows, and zeroes everywhere else. Similarly, we may enter ones on the main diagonal of the submatrix of $\beta^{\prime}$ that excludes the last two columns.

With two fundamental trends, no irreducible cointegrating relation can involve more than three variables. Exhaustive consideration along Davidson's lines would produce 21 possible cointegrating pairs and 35 possible cointegrating triples. Similarly, we need to consider possible weak exogeneity of variables within each irreducibly cointegrating subset. Most of subsets are not irreducibly cointegrating or do not contain weakly exogenous variables, so rather than tediously listing the weak-exogeneity status of all 56 subsets systematically, we just note the salient ones.

From the facts that $\mathrm{CI}\left(\left\{A, T_{1}\right\}\right)$ and that there are no other variables in this cointegration group and that $T_{1} \mapsto A$, Rule 2 implies $T_{1} \rightarrow \mathrm{~A}$, which justifies the placement of $\beta_{A T_{1}}$ in row 1 of $\beta^{\prime}$ and zeroes in the remaining unassigned places in that row. Analogous reasoning with respect to $\left\{C, T_{2}\right\}$ implies $T_{2} \rightarrow$ $C$ and justifies the placement of $\beta_{C T_{2}}$ and the zeroes in row 3. Again, with respect to $\{B, D\}$, analogous reasoning justifies the placement of $\beta_{D B}$ and the zeroes in row 4 . In addition, in this case, Rule 1 and the fact that $B \mapsto D$ imply that $\operatorname{not}(D \rightarrow B)$ and justify the zero in row 2 , column 4 .

Rule 3 and the facts that $\operatorname{IC}\left(\left\{T_{1}, T_{2}, B\right\}\right)$, that $B$ is not part of a cointegration group with either $T_{1}$ or $T_{2}$, and that $\left\{T_{1}, T_{2}\right\} \mapsto B$ allows us to identify the collider $T_{1} \rightarrow B \leftarrow T_{2}$ and justifies the placement of $\beta_{B T_{1}}$ and $\beta_{B T_{2}}$ and the remaining zeroes in row 2 of $\beta^{\prime}$.

Rules 3 and the facts that $\operatorname{IC}(\{B, C, E\}),\left(\{B, C\} \mapsto E\right.$, and not $\left(C \mapsto T_{2}\right)$, with which it forms a cointegration group, allows us to identify the collider $B \rightarrow E \leftarrow C$ and justifies the placement of $\beta_{E B}$ and $\beta_{E C}$ and the zeroes in row 5 of $\beta^{\prime}$. With that, we were able to recover the entire DGP graph using only the facts of cointegration and weak exogeneity.

### 5.2. Long-Run Causal Search in the Presence of Latent Trends

The CVARs typically estimated in practice most often do not contain variables that are weakly exogenous for the whole system, which could, therefore, be identified as the conduit of the fundamental
trends to the other variables in the system. It is, therefore, worth considering how the principles of search might operate when fundamental trends are latent variables. It is possible to apply the rules of Section 4.2 to the variables generated according to Equation (7) when only the ordinary variables $(A, B$, $C, D, E)$, but not the fundamental trends $\left(T_{1}\right.$ and $\left.T_{2}\right)$, are observed.

For some of the causal edges, the reasoning of Section 4.3 is still applicable, and we would be able to infer the edges shown in Figure 7: B $\rightarrow$ and $B \rightarrow \mathrm{E} \leftarrow \mathrm{C}$. The remainder of Figure 7 requires further comment.


Figure 7. Recoverable structure of Figure 2 graph when fundamental trends are latent.
We are unable to infer the edges between $T_{1}, T_{2}$ and $A, B$, and $C$ for the simple reason that the two fundamental trends are not observed and the inference of the edges in which they are involved requires their observability. However, we do know from the fact that the cointegration rank is 3 that there are two fundamental trends. What we cannot say, however, is precisely how those two trends enter directly into the observable system. They may, in fact, be transmitted through ordinary variables that are also latent. We do know that they must enter through $A, B$, or $C$. If that were not the case and a fundamental trend entered through $D$ or $E$, we would not have found that $\mathrm{CI}(\{B, D\})$ or $\{B, C\} \mapsto E$. This is indicated in Figure 7 by the oval enclosing the ordinary variables and the circles (indicating their latency) around the fundamental trends. The arrows running from the latent fundamental trends to the oval, stopping short of the particular variables indicates that we know that these variables are caused by these trends, albeit we do not know exactly what the connections are. Thus, instead of (7), we can fill in the causally ordered CVAR Equation (15) with the ambiguous information depicted in Figure 7, where the question marks indicate parameters that correspond to possible, but yet-to-be-determined causal edges.

$$
\begin{align*}
& \Delta \xi_{\mathrm{t}}=\boldsymbol{\Psi} \xi_{\mathrm{t}-1}+\mathbf{H}_{\mathrm{t}} \\
& \quad=\alpha \beta^{\prime} \xi_{\mathrm{t}-1}+\mathbf{H}_{\mathrm{t}}=\left[\begin{array}{ccccc}
\alpha_{\mathrm{AA}} & 0 & 0 & 0 & 0 \\
0 & \alpha_{\mathrm{BB}} & 0 & 0 & 0 \\
0 & 0 & \alpha_{\mathrm{CC}} & 0 & 0 \\
0 & 0 & 0 & \alpha_{\mathrm{DD}} & 0 \\
0 & 0 & 0 & 0 & \alpha_{\mathrm{EE}} \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{array}\right]\left[\begin{array}{ccccccc}
1 & ? & ? & 0 & 0 & ? & ? \\
? & 1 & ? & 0 & 0 & ? & ? \\
? & ? & 1 & 0 & 0 & ? & ? \\
0 & \beta_{\mathrm{DB}} & 0 & 1 & 0 & 0 & 0 \\
0 & \beta_{\mathrm{EB}} & \beta_{\mathrm{EC}} & 0 & 1 & 0 & 0
\end{array}\right]\left[\begin{array}{c}
\mathrm{A} \\
\mathrm{~B} \\
\mathrm{C} \\
\mathrm{D} \\
\mathrm{E} \\
\mathrm{~T}_{1} \\
\mathrm{~T}_{2}
\end{array}\right]_{\mathrm{t}-1}+\mathbf{H}_{\mathrm{t}} \tag{15}
\end{align*}
$$

Equation (15) depicts what observables imply about the DGP and not just facts about the observables themselves. Here the two trends are not observable, but we know that there are two latent trends because none of the observable variables is weakly exogenous when one considers the whole set of observable variables, which again justifies the placement of the two zero rows in $\alpha$.

Neither the graph nor (15) conveys all the information that we have. We know, for instance, that there are two fundamental trends and that at least one of the fundamental trends must causally influence each of $A, B$, and $C$. If that were not so, then the only way that all three variables could carry the trends and be irreducibly cointegrated would be for them to form a collider group in which one pair is weakly exogenous for the remaining variable. Given the DGP, we know that the weak exogeneity search should not find such a pattern. Furthermore, we know that no two of $A, B$, and $C$
could have a common latent cause. If that were not true, that pair would form a cointegration group, which, given the DGP, the search for cointegrating pairs should not find such a cointegration group. These two conclusions imply that each of the three observed variables carries the fundamental trends in distinct proportions. These facts place restrictions on how the last two columns of the $\beta^{\prime}$ in (15) can be filled in to be consistent with the DGP. In particular, in $3 \times 2$ submatrix in the upper right-hand corner of $\beta^{\prime}$, at least one row must contain two nonzero entries and the remaining two rows cannot have zeroes in the same column. This guarantees that the variables $A, B, C$ form a cointegration group without also forming a collider group with weakly exogenous parents.

## 6. Conclusions

In the history of econometrics, the problem of identification and the notion of causal order have long been connected-both in the work of Simon and the early Cowles Commission program and in the literature on SVARs. Typically, economists have relied heavily on the idea that a priori restrictions derived somehow from economic theory would provide the needed identification. Recent work on graphical causal modeling, however, has shown that there is often unexploited information that could provide a firmer, empirical basis for identification. In the case of cross-sectional data or the contemporaneous causal orderings of SVARs, the graphical causal modelers have stressed the information contained in conditional independence relationship encoded in the probability distribution of the data. Conditional independence may also be a resource in the case of the long-run dynamics of the CVAR, although the fact that nonstationary data involves non-standard distributions poses some challenges. We have suggested here that nonstationary data also present the opportunity to take a different approach.

Where do the trends we observe among macroeconomic variables come from? We showed that it is possible for the structure of the DGP to be such that a set of observable variables trends without any fundamental trends acting as drivers. Yet, we have argued that these cases rely on particular configurations of coefficients that are likely not to be robust to small changes in coefficients and that call out for an economic explanation of why they arise at all. Once a distinction is drawn between fundamental trends and ordinary variables, it is clear that a more robust account for nonstationary behavior is that it is transmitted from its fundamental sources to variables that without these fundamental trends as direct or indirect causes would not naturally be nonstationary. In typical CVAR analysis, econometricians mostly do not find variables that themselves can be identified as the source of fundamental trends. This suggests that, in most cases, fundamental trends are latent variables, and any sort of structural or causal analysis of CVARs must account for their latency.

We suggested-somewhat informally-that combining Davidson's suggestion of a comprehensive search for sets of irreducible cointegrating relations with a similar comprehensive search of weak exogeneity among those sets could provide a non-a priori empirical basis for discovering identifying restrictions on cointegrating relations, as well as information on causal direction. We showed that in a simple example, the complete causal graph of the CVAR could be recovered. But, in most cases in the face of latent variables, these restrictions are unlikely to provide complete identification. Nevertheless, as in our illustration, some of the cointegrating relations may be identified, even when there are latent trends. It is also possible that, in some cases, it would be possible to recover estimates of the trends using state-space methods (see, e.g., Johansen and Tabor 2017). Finally, viewing the CVAR through the lens of latent fundamental trends reinforces Juselius's advocacy of simple-to-general modeling in the CVAR context (Juselius 2006, Chapter 22, especially. Sections 22.2.3 and 22.3). Cointegrating relations are robust to widening the data set to include more variables. The aim of such widening can be seen as an effort to discover the observable variables that are the counterpart of the latent trends in narrower data sets.

Funding: This research received no external funding.
Conflicts of Interest: The author declares no conflict of interest.

## References

Cooley, Thomas F., and Stephen F. LeRoy. 1985. Atheoretical Macroeconomics: A Critique. Journal of Monetary Economics 16: 283-308. [CrossRef]
Cooper, Gregory F. 1999. An Overview of the Representation and Discovery of Causal Relationships using Bayesian Networks. In Computation, Causation, and Discovery. Edited by Clark Glymour and Gregory F. Cooper. Menlo Park: American Association for Artificial Intelligence, Cambridge: MIT Press, pp. 3-64.
Davidson, James, and Stephen Hall. 1991. Cointegration in Recursive Systems. Economic Journal 101: 109-251. [CrossRef]
Davidson, James. 1998. Structural Relations, Cointegration and Identification: Som Simple Results and Their Application. Journal of Econometrics 87: 87-113. [CrossRef]
Demiralp, Selva, and Kevin D. Hoover. 2003. Searching for the Causal Structure of a Vector Autoregression. Oxford Bulletin of Economics and Statistics 65: 745-67. [CrossRef]
Demiralp, Selva, Kevin D. Hoover, and Stephen J. Perez. 2008. A Bootstrap Method for Identifying and Evaluating a Structural Vector Autoregression. Oxford Bulletin of Economics and Statistics 70: 509-33. [CrossRef]
Duarte, Pedro Garcia, and Kevin D. Hoover. 2012. Observing Shocks. In Histories of Observation in Economics. (Supplement to History of Political Economy 44). Edited by Harro Maas and Mary Morgan. Durham: Duke University Press, pp. 226-49.
Engle, Robert F., and Clive W. J. Granger. 1987. Cointegration and Error Correction Representation: Estimation and Testing. Econometrica 55: 251-76. [CrossRef]
Engle, Robert F., David F. Hendry, and Jean-François Richard. 1983. Exogeneity. Econometrica 51: 277-304. [CrossRef]
Granger, Clive W. J. 1969. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica 37: 424-38. [CrossRef]
Granger, Clive W. J. 1995. Commentary [on Stephen F. LeRoy, Causal Orderings]. In Macroeconometrics: Developments, Tensions, and Prospects. Edited by Kevin D. Hoover. Boston: Kluwer, pp. 229-352.
Haavelmo, Trygve. 1944. The Probability Approach in Econometrics. Econometrica 12: iii-115. [CrossRef]
Hood, William C., and Tjalling Koopmans, eds. 1953. Studies in Econometric Method. (Cowles Commission Monograph vol. 14). New York: Wiley.
Hoover, Kevin D. 1990. The Logic of Causal Inference: Econometrics and the Conditional Analysis of Causation. Economics and Philosophy 6: 207-34. [CrossRef]
Hoover, Kevin D. 2001. Causality in Macroeconomics. Cambridge: Cambridge University Press.
Hoover, Kevin D. 2004. Lost Causes. Journal of the History of Economic Thought 26: 149-64. [CrossRef]
Hoover, Kevin D. 2008. Causality in Economics and Econometrics. In The New Palgrave Dictionary of Economics, 2nd ed. Edited by Steven N. Durlauf and Lawrence E. Blume. London: Palgrave Macmillan. [CrossRef]
Hoover, Kevin D. 2012. Economic Theory and Causal Inference. In Handbook of the Philosophy of Economics. volume 13 of Dov Gabbay, Paul Thagard, and John Woods, general editors. Edited by Uskali Mäki. Amsterdam: Elsevier/North-Holland, pp. 89-113.
Hoover, Kevin D. 2014. On the Reception of Haavelmo's Econometric Thought. Journal of the History of Economic Thought 36: 45-65. [CrossRef]
Hoover, Kevin D. 2015. The Ontological Status of Shocks and Trends in Macroeconomics. Synthese, 3509-32. [CrossRef]
Hoover, Kevin D., and Katarina Juselius. 2015. Trygve Haavelmo's Experimental Methodology and Scenario Analysis in a Cointegrated Vector Autoregression. Econometric Theory 31: 249-74. [CrossRef]
Hoover, Kevin D., and Òscar Jordá. 2001. Measuring Systematic Monetary Policy. Federal Reserve Bank of St. Louis Review 83: 113-37. [CrossRef]
Hoover, Kevin D., Søren Johansen, and Katarina Juselius. 2008. Allowing the Data to Speak Freely: The Macroeconometrics of the Cointegrated Vector Autoregression. American Economic Review: Papers and Proceedings 98: 251-55. [CrossRef]
Johansen, Søren, and Katarina Juselius. 2014. An Asymptotic Invariance Property of the Common Trends Under Linear Transformation of the Data. Journal of Econometrics 178: 310-15. [CrossRef]
Johansen, Søren, and Morten Nyboe Tabor. 2017. Cointegration between Trends and Their Estimators in State Space Models and Cointegrated Vector Autoregressive Models. Econometrics 5: 36. [CrossRef]

Johansen, Søren. 1995. Likelihood-based Inference in Cointegrated Vector Autoregressive Models. Oxford: Oxford University Press.
Johansen, Søren. 2019. Cointegration and Adjustment in the CVAR $(\infty)$ Representation of Some Partially Observed CVAR(1) Models. Econometrics 7: 2. [CrossRef]
Juselius, Katarina. 2006. The Cointegrated VAR Model. Oxford: Oxford University Press.
Koopmans, Tjalling C., ed. 1950. Statistical Inference in Dynamic Economic Models. (Cowles Commission Monograph 10). New York: Wiley.

LeRoy, Stephen F. 1995. Causal Orderings. In Macroeconometrics: Developments, Tensions, and Prospects. Edited by Kevin D. Hoover. Boston: Kluwer, pp. 235-52.
Liu, Ta-Chung. 1960. Underidentification, Structural Estimation, and Forecasting. Econometrica 28: 855-65. [CrossRef]
Morgan, Mary S. 1990. The History of Econometric Ideas. Cambridge: Cambridge University Press.
Pagan, A. R., and M. Hashem Pesaran. 2008. Econometric analysis of structural systems with permanent and transitory shocks. Journal of Economic Dynamics and Control 32: 3376-95. [CrossRef]
Pearl, Judea. 2009. Causality: Models, Reasoning, and Inference, 2nd ed. Cambridge: Cambridge University Press.
Pesaran, M. Hashem, and Ron P. Smith. 1998. Structural Analysis of Cointegrating VARs. Journal of Economic Surveys 12: 471-506. [CrossRef]
Pesaran, M. Hashem, and Yougcheoi Shin. 2002. Long-run Structural Modelling. Econometric Reviews 21: 49-87. [CrossRef]
Phiromswad, Piyachart, and Kevin D. Hoover. 2013. Selecting Instrumental Variables: A Graph-Theoretic Approach. August 28. Available online: https://ssrn.com/abstract=2318552 (accessed on 28 July 2020).
Reichenbach, Hans. 1956. The Direction of Time. Berkeley: University of California Press.
Richardson, Thomas. 1996. A Discovery Algorithm for Directed Cyclic Graphs. In Uncertainty in Artificial Intelligence: Proceedings of the Twelfth Conference. Edited by Erich Horvitz and Finn Jensen. San Francisco: Morgan Kaufmann, pp. 454-61.
Simon, Herbert A. 1953. Causal Order and Identifiability. In Hood and Koopmans. New York: Wiley, vol. 14, pp. 49-74.
Sims, Christopher A. 1972. Money, Income and Causality. American Economic Review 62: 540-52.
Sims, Christopher A. 1980. Macroeconomics and Reality. Econometrica 48: 1-48. [CrossRef]
Sims, Christopher A. 1982. Policy Analysis with Econometric Models. Brookings Papers on Economic Activity 13: 107-52. [CrossRef]
Sims, Christopher A. 1986. Are Forecasting Models Usable for Policy Analysis? Federal Reserve Bank of Minneapolis Quarterly Review 10: 2-15. [CrossRef]
Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. Causation, Prediction, and Search, 2nd ed. Cambridge: MIT Press.
Swanson, Norman R., and Clive W. J. Granger. 1997. Impulse Response Functions Based on a Causal Approach to Residual Orthogonalization in Vector Autoregressions. Journal of the American Statistical Association 92: 357-67. [CrossRef]
White, Halbert, and Davide Pettenuzzo. 2014. Granger-causality, Exogeneity, and Economic Policy Analysis. Journal of Econometrics 178: 316-30. [CrossRef]
White, Halbert, and Karim Chalak. 2009. Settable Systems: An Extension of Pearl's Causal Model with Optimization, Equilibrium, and Learning. Journal of Machine Learning Research 10: 1759-99.
White, Halbert, and Xun Lu. 2010. Granger-causality and Dynamic Structural Systems. Journal of Financial Econometrics 8: 193-243. [CrossRef]
Wold, Hermann. 1960. A Generalization of Causal Chain Models. Econometrica 28: 277-326. [CrossRef]
© 2020 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

## Article

# Evaluating Forecasts, Narratives and Policy Using a Test of Invariance 

Jennifer L. Castle ${ }^{1}$, David F. Hendry ${ }^{2, *}$ and Andrew B. Martinez ${ }^{2}$<br>1 Magdalen College and Institute for New Economic Thinking, Oxford Martin School, University of Oxford, OX1 4AU Oxford, UK; jennifer.castle@magd.ox.ac.uk<br>2 Department of Economics, and Institute for New Economic Thinking, Oxford Martin School, University of Oxford, OX1 3UQ Oxford, UK; andrew.martinez@economics.ox.ac.uk<br>* Correspondence: david.hendry@nuffield.ox.ac.uk; Tel.: +44-1865-278654<br>Academic Editors: Rocco Mosconi and Paolo Paruolo<br>Received: 17 December 2016; Accepted: 23 August 2017; Published: 1 September 2017


#### Abstract

Economic policy agencies produce forecasts with accompanying narratives, and base policy changes on the resulting anticipated developments in the target variables. Systematic forecast failure, defined as large, persistent deviations of the outturns from the numerical forecasts, can make the associated narrative false, which would in turn question the validity of the entailed policy implementation. We establish when systematic forecast failure entails failure of the accompanying narrative, which we call forediction failure, and when that in turn implies policy invalidity. Most policy regime changes involve location shifts, which can induce forediction failure unless the policy variable is super exogenous in the policy model. We propose a step-indicator saturation test to check in advance for invariance to policy changes. Systematic forecast failure, or a lack of invariance, previously justified by narratives reveals such stories to be economic fiction.


Keywords: forediction; invariance; super exogeneity; indicator saturation; co-breaking; Autometrics

JEL Classification: C22; C51

## 1. Introduction

The Bank of England's quarterly Inflation Reports announces its projections of CPI inflation and 4-quarter real GDP growth for the next two years. For example, those made in November 2009 are shown in Figure 1. Accompanying these forecast distributions are textual explanations for the forecasts, an excerpt from which is: ${ }^{1}$

CPI inflation looked set to rise sharply in the near term. Further out, downward pressure from the persistent margin of spare capacity was likely to bear down on inflation for some time to come.

In his speech on the Report, the Governor stressed:
It is more likely than not that later this year I will need to write a letter to the Chancellor to explain why inflation has fallen more than 1 percentage point below the target (of $2 \%$ ). The stimulus to demand, combined with a turnaround in the stock cycle and the effects of the depreciation in sterling, is likely to drive a recovery in activity.

[^53]We term the published numerical forecast a 'direct forecast', whereas one constructed from the narrative, as in e.g., Ericsson (2016) for the USA, is called a 'derived forecast'. Taken together, we call the joint production of the numerical forecast and the accompanying narrative a 'forediction', intended to convey a forecast made alongside a story (diction) that describes the forecast verbally. ${ }^{2}$ In this paper, we investigate whether a close link between the direct and derived forecasts sustains an evaluation of the resulting foredictions and their associated policies.

Figure 1 shows the direct forecasts for CPI inflation and 4-quarter real GDP growth from the Bank of England November 2009 Inflation Report, with the outturns in October 2011. The outturn for CPI inflation 2 years out lay well above the central $90 \%$ of the distribution for the projected probabilities of CPI inflation outturns, well above the $2 \%$ target, yet the outturn for GDP lay in the lowest band. The large forecast error on the direct CPI inflation forecast refutes the above narrative, and would do so more generally when the derived forecast is closely similar, resulting in a forediction failure. The next step would be to investigate the validity of policy decisions made on the basis of the forediction, namely here, the Bank of England decided to hold the Bank Rate at $0.5 \%$ and finance a further $£ 25$ billion of asset purchases within 3 months. ${ }^{3}$


Figure 1. CPI inflation and 4-quarter real GDP growth forecasts, based on market interest rate expectations and $£ 200$ billion asset purchases: November 2009 Bank of England Inflation Report, with October 2011 outturns.

The two objectives of this paper are (i) establishing when systematic forecast failure, defined as large, persistent deviations of the outturns from the forecasts, entails forediction failure and when that in turn implies policy invalidity; and (ii) whether failures of invariance in policy models are detectable in advance, which should enable more robust policy models to be developed.

There are four key steps in the process of evaluating foredictions and associated policies. The first is to establish whether the direct and derived forecasts are almost the same. Then forecasts are indeed closely described by the accompanying narrative in the sense formalized by Ericsson (2016), namely accurate estimates of the forecasts can be derived by quantifying the narrative. For example, regressing the derived forecasts on the direct should deliver a highly significant relation with a coefficient near unity. The second step involves testing whether systematic forecast failure of the direct forecasts occurs, for which many tests exist. If the direct and derived forecasts are closely

[^54]linked, as checked in the first step, then systematic forecast failure of the direct forecasts would imply systematic forecast failure of the derived forecasts. This is forediction failure, checked by testing (a) if the narrative-derived forecasts remain linked to the direct forecasts, and (b) they also differ significantly from the realized outcomes. In that case, the accompanying narrative must also be rejected. In the third step, if a policy implementation had been justified by the forediction, then it too must be invalid after forediction failure. This is harder to quantify unless there is a known policy rule, such as a Taylor rule linking interest rate responses to inflation, but the policy link may be stated explicitly in the narrative. The fourth step is to test if the forediction failure actually resulted from the policy change itself because of a lack of invariance of the model used in the policy analysis to that change. In that case, the policy model is also shown to be invalid.

Evaluating the validity of policy analysis involves two intermediate links. First, genuine causality from policy instruments to target variables is essential for policy effectiveness if changing the policy variable is to affect the target: see e.g., Cartwright (1989). That influence could be indirect, as when an interest rate affects aggregate demand which thereby changes the rate of inflation. Secondly, the invariance of the parameters of the empirical models used for policy to shifts in the distributions of their explanatory variables is also essential for analyses to correctly represent the likely outcomes of policy changes. Thus, a policy model must not only embody genuine causal links in the data generation process (DGP), invariance also requires the absence of links between target parameters and policy-instrument parameters, since the former cannot be invariant to changes in the latter if their DGP parameters are linked, independently of what modellers may assume: see e.g., Hendry (2004) and Zhang et al. (2015). Consequently, weak exogeneity matters in a policy context because its two key requirements are that the relevant model captures the parameters of interest for policy analyses, and that the parameters of the policy instrument and the target processes are unconnected in the DGP, not merely in their assumed parameter spaces. Since changes in policy are involved, valid policy then requires super exogeneity of the policy variables in the target processes: see Engle et al. (1983).

To this end, the paper proposes a step-indicator saturation test to check in advance for invariance to policy changes. Step-indicator saturation (SIS) is designed to detect location shifts of unknown magnitude, location and frequency at any point in the sample, see Castle et al. (2015). Super exogeneity, which is required for policy validity, can be tested by checking whether any significant location shifts in models of the policy instruments are significant in the model of the target variables. The proposed test is an extension of the test for super exogeneity using impulse indicator saturation proposed in Hendry and Santos (2010).

Combining these ideas, forediction evaluation becomes a feasible approach to checking the validity of policy decisions based on forecasts linked to a narrative that the policy agency wishes to convey. Policy decisions can fail because empirical models are inadequate, causal links do not exist, parameters are not invariant, the story is incorrect, or unanticipated events occur. By checking the properties of the direct and derived forecasts and the invariance of the policy model to the policy change envisaged, the source of forediction failure may be more clearly discerned, hopefully leading to fewer bad mistakes in the future.

The structure of the paper is as follows. Section 2 explores the link between direct and derived forecasts with accompanying narratives that are often used to justify policy decisions, where Section 2.1 discusses whether narratives or forecasts come first, or are jointly decided. Section 3 considers what can be learned from systematic forecast failure, and whether there are any entailed implications for forediction failure. More formally, Section 3.1 outlines a simple theoretical policy DGP; Section 3.2 describes what can be learned from systematic forecast failure using a taxonomy of forecast errors for a mis-specified model of that DGP; Section 3.3 provides a numerical illustration; and Section 3.4 notes some of the implications of forediction failure for economic theories using inter-temporal optimization. Parametric time-varying models are not explicitly considered, but Section 6.1 investigates an approach that potentially allows coefficients to change in many periods; our analysis would extend to handling location shifts in models with time-varying coefficients but constant underlying parameters, as in
structural time series models which have constant-parameter ARIMA representations. Section 4 presents a two-stage test for invariance using automatic model selection to implement step-indicator saturation, which extends impulse-indicator saturation (IIS): see Hendry et al. (2008). Section 5 reports some simulation evidence on its first-stage null retention frequency of irrelevant indicators (gauge) and its retention frequency of relevant indicators (potency), as well as its second-stage rejection frequencies in the presence or absence of invariance: see Johansen and Nielsen (2016) for an analysis of gauge in IIS. Section 6 applies this SIS-based test to the small artificial-data policy model analyzed in Section 3.1, and Section 6.1 investigates whether a shift in a policy parameter can be detected using multiplicative indicator saturation (MIS), where step indicators are interacted with variables. Section 7 summarizes the forecast error taxonomy and associated relevant tests, and Section 8 considers how intercept corrections can improve forecasts without changing the forecasting model's policy responses. Section 9 concludes.

## 2. Forediction: Linking Forecasts, Narratives, and Policies

Links between forecasts and their accompanying narratives have been brought into new salience through innovative research by Stekler and Symington (2016) and Ericsson (2016). The former develop quantitative indexes of optimism/pessimism about the US economy by analyzing the qualitative information in the minutes from Federal Open Market Committee (FOMC) meetings, scaled to match real GDP growth rates in percentages per annum. The latter calibrates the Stekler and Symington (2016) index for real GDP growth, denoted FMI, to past GDP growth outcomes, removing its truncation to $(-1,+4)$, and shows that the resulting index can provide excellent post-casts of the Fed's 'Greenbook' forecasts of 2006-10 (which are only released after a 5-year delay). Clements and Reade (2016) consider whether Bank of England narratives provide additional information beyond their inflation forecasts.

For many years, Central Banks have published narratives both to describe and interpret their forecasts, and often justify entailed policies. Important examples include the minutes from Federal Open Market Committee (FOMC) meetings, the Inflation Reports of the Bank of England, and International Monetary Fund (IMF) Reports. For Banca d'Italia, Siviero and Terlizzese (2001) state that:
...forecasting does not simply amount to producing a set of figures: rather, it aims at assembling a fully-fledged view-one may call it a "story behind the figures"—of what could happen: a story that has to be internally consistent, whose logical plausibility can be assessed, whose structure is sufficiently articulated to allow one to make a systematic comparison with the wealth of information that accumulates as time goes by.

Such an approach is what we term forediction: any claims as to its success need to be evaluated in the light of the widespread forecast failure precipitated by the 2008-9 Financial Crisis. As we show below, closely tying narratives and forecasts may actually achieve the opposite of what those authors seem to infer, by rejecting both the narratives and associated policies when forecasts go wrong.

Sufficiently close links between direct forecasts and forecasts derived from their accompanying narratives, as highlighted by Stekler and Symington (2016) and Ericsson (2016), entail that systematic forecast failure vitiates any related narrative and its associated policy. This holds irrespective of whether the direct forecasts lead to the narrative, or the forecasts are modified (e.g., by 'judgemental adjustments') to satisfy a preconceived view of the future expressed in a narrative deliberately designed to justify a policy, or the two are iteratively adjusted as in the above quote, perhaps to take account of informal information available to a panel such as the Bank of England's Monetary Policy Committee (MPC). In all three cases, if the direct and derived forecasts are almost the same, and the narratives reflect cognitive models or theories, then the large forecast errors around the 'Great Recession' would also refute such thinking, as addressed in Section 3. However, when direct and derived forecasts are not tightly linked, failure in one need not entail failure in the other, but both can be tested empirically.

### 2.1. Do Narratives or Forecasts Come First?

Forcing internal consistency between forecasts and narratives, as both Siviero and Terlizzese (2001) and Pagan (2003) stress, could be achieved by deciding the story, then choosing add factors to achieve it, or vice versa, or by a combination of adjustments. The former authors appear to suggest the third was common at Banca d'Italia, and the fact that Bank of England forecasts are those of the MPC based on the information it has, which includes forecasts from the Bank's models, suggests mutual adjustments of forecasts and narratives, as in e.g., (Bank of England 2015, p. 33) (our italics):

The projections for growth and inflation are underpinned by four key judgements.
Not only could add factors be used to match forecasts to a narrative, if policy makers had a suite of forecasting models at their disposal, then the weightings on different models could be adjusted to match their pooled forecast to the narrative, or both could be modified in an iterative process. Genberg and Martinez (2014) show the link between narratives and forecasts at the International Monetary Fund (IMF), where forecasts are generated on a continuous basis through the use of a spreadsheet framework that is occasionally supplemented by satellite models, as described in Independent Evaluation Office (2014). Such forecasts 'form the basis of the analysis [...] and of the [IMF's] view of the outlook for the world economy' (p.1). Thus, adjustments to forecasts and changes to the associated narratives tend to go hand-in-hand at many major institutions.

In a setting with several forecasting models used by an agency that nevertheless delivered a unique policy prescription, possibly based on a 'pooled forecast', then to avoid implementing policy incorrectly, super exogeneity with respect to the policy instrument would seem essential for every model used in that forecast. In practice, the above quotes suggest some Central Banks act as though there is a unique policy model constrained to match their narrative, where all the models in the forecasting suite used in the policy decision are assumed to be invariant to any resulting changes in policy. This requirement also applies to scenario studies as the parameters of the models being used must be valid across all the states examined. Simply asserting that the policy model is 'structural' because it is derived from a theory is hardly sufficient. Akram and Nymoen (2009) consider the role of empirical validity in monetary policy, and caution against over-riding it. Even though policy makers recognise that their policy model may not be the best forecasting model, as illustrated by the claimed trade-off between theory consistency and empirical coherence in Pagan (2003), forecast failure can still entail forediction failure, as shown in Section 3.

### 2.2. Is There a Link between Forecasts and Policy?

Evaluating policy-based forediction failure also requires there to be a link between the forecasts and policy changes. This can either occur indirectly through the narratives or directly through the forecasts themselves (e.g., through a forward looking Taylor rule). That direct forecasts and forecasts derived from the narratives are closely linked implies that policies justified by those narratives, are also linked to the forecasts.

In their analysis of the differences between the Greenbook and FOMC forecasts, Romer and Romer (2008) find that there is a statistically significant link between differences in these two forecasts and monetary policy shocks. This can be interpreted as evidence suggesting that policy makers forecasts influence their decisions above and beyond other informational inputs. As such, the policy decision is both influenced by and influences the forecasts.

Regardless of this influence, the link between forecasts and policy depends on how the forecasts are used. Ellison and Sargent (2012) illustrate that it is possible for 'bad forecasters' to be 'good policymakers'. In this sense, forecast failure could potentially be associated with policy success. However, this is less likely when focusing on systematic forecast failure. Sinclair et al. (2016) find that forecast failure at the Fed is associated with policy failure. They argue that in 2007-08 the Greenbook's overpredictions of output and inflation "resulted in a large contractionary [monetary policy] shock" of around 7 percentage points. By their calculations, using a forward-looking

Taylor rule, this forecast-error-induced shock reduced real growth by almost 1.5 percentage points. Furthermore, Independent Evaluation Office (2014) illustrates that in IMF Programs, the forecasts are often interpretable as the negotiated policy targets. This suggests that forecasts, narratives and policy are sufficiently closely linked so that systematic forediction failure can be used to refute the validity of policy decisions.

## 3. Forecast Failure and Forediction Failure

Systematic forecast failure by itself is not a sufficient condition for rejecting an underlying theory or any associated forecasting model, see, e.g., Castle and Hendry (2011). Furthermore, forecasting success may, but need not, 'corroborate' the forecasting model and its supporting theory: see Clements and Hendry (2005). Indeed, Hendry (2006) demonstrates a robust forecasting device that can outperform the forecasts from an estimated in-sample DGP after a location shift. Nevertheless, when several rival explanations exist, forecast failure can play an important role in distinguishing between them as discussed by Spanos (2007). Moreover, systematic forecast failure almost always rejects any narrative associated with the failed forecasts and any policy implications therefrom, so inevitably results in forediction failure.

### 3.1. A Simple Policy Data Generation Process

In this section, we develop a simple theoretical policy model which is mis-specified for its economy's DGP. At time $T$, the DGP shifts unexpectedly, so there is forecast failure. However, even if the DGP did not shift, because the policy model is mis-specified, a policy change itself can induce forecast failure. In our model, the policy agency decides on an action at precisely time $T$, so the two shifts interact. We provide a taxonomy of the resulting sources of failure, and what can be inferred from each component.

Let $y_{t+1}$ be a policy target, say inflation, $z_{t}$ the instrument an agency controls to influence that target, say interest rates, where $x_{t+1}$ is the variable that is directly influenced by the instrument, say aggregate excess demand, which in turn directly affects inflation. Also, let $w_{1, t}$ represent the net effect of other variables on the target, say world inflation in the domestic currency, and $w_{2, t}$ denote additional forces directly affecting domestic demand, where the $\left\{w_{i, t}\right\}$ are super exogenous for the parameters in the DGP. The system is formalized as:

$$
\begin{equation*}
y_{t}=\gamma_{0}+\gamma_{1} x_{t}+\gamma_{2} w_{1, t}+\epsilon_{t} \tag{1}
\end{equation*}
$$

where for simplicity $\epsilon_{t} \sim \operatorname{IN}\left[0, \sigma_{\epsilon}^{2}\right]$ with $\gamma_{1}>0$ and $\gamma_{2}>0$; and :

$$
\begin{equation*}
x_{t}=\beta_{0}+\beta_{1} z_{t-1}+\beta_{2} w_{2, t}+v_{t} \tag{2}
\end{equation*}
$$

with $v_{t} \sim \operatorname{IN}\left[0, \sigma_{v}^{2}\right]$ where $\beta_{1}<0$ and $\beta_{2}>0$. In the next period, solving as a function of policy and exogenous variables:

$$
\begin{equation*}
y_{t+1}=\left(\gamma_{0}+\gamma_{1} \beta_{0}\right)+\gamma_{1} \beta_{1} z_{t}+\gamma_{2} w_{1, t+1}+\gamma_{1} \beta_{2} w_{2, t+1}+\left(\epsilon_{t+1}+\gamma_{1} v_{t+1}\right) \tag{3}
\end{equation*}
$$

Consequently, a policy shift moving $z_{t}$ to $z_{t}+\delta$ would, on average and ceteris paribus, move $y_{t+1}$ to $y_{t+1}+\delta \gamma_{1} \beta_{1}$. We omit endogenous dynamics to focus on the issues around forediction.

In a high-dimensional, wide-sense non-stationary world, where the model is not the DGP, and parameters need to be estimated from data that are not accurately measured, the policy agency's model of inflation is the mis-specified representation:

$$
\begin{equation*}
y_{t}=\lambda_{0}+\lambda_{1} x_{t}+e_{t} \tag{4}
\end{equation*}
$$

where $\lambda_{0}$ and $\lambda_{1}$ are obtained as the OLS estimates of the regression of $y_{t}$ on $x_{t}$, thus omitting $w_{1, t}$. Also, its model of demand is mis-specified as:

$$
\begin{equation*}
x_{t}=\theta_{0}+\theta_{1} z_{t-1}+u_{t} \tag{5}
\end{equation*}
$$

where $\theta_{0}$ and $\theta_{1}$ are obtained as the OLS estimates of the regression of $x_{t}$ on $z_{t}$, thus omitting $w_{2, t}$. This system leads to its view of the policy impact of $z_{t}$ at $t+1$ as being on average and ceteris paribus:

$$
\begin{equation*}
y_{t+1 \mid t}=\left(\lambda_{0}+\lambda_{1} \theta_{0}\right)+\lambda_{1} \theta_{1} z_{t}+\left(e_{t+1}+\lambda_{1} u_{t+1}\right) \tag{6}
\end{equation*}
$$

so a policy shift moving $z_{t}$ to $z_{t}+\delta$ would be expected to move $y_{t+1 \mid t}$ to $y_{t+1 \mid t}+\delta \widehat{\lambda}_{1} \widehat{\theta}_{1}$ where $\widehat{\lambda}_{1}$ and $\widehat{\theta}_{1}$ are in-sample estimates of $\lambda_{1}$ and $\theta_{1}$ respectively. If the agency wishes to lower inflation when $\lambda_{1}>0$ and $\theta_{1}<0$, it must set $\delta>0$ such that $\delta \widehat{\lambda}_{1} \widehat{\theta}_{1}<0$ (e.g., -0.01 corresponds to a 1 percentage point reduction in the annual inflation rate).

The equilibrium means in the stationary world before any shifts or policy changes are:

$$
\mathrm{E}\left[y_{t}\right]=\gamma_{0}+\gamma_{1} \beta_{0}+\gamma_{1} \beta_{1} \mu_{z}+\gamma_{2} \mu_{w_{1}}+\gamma_{1} \beta_{2} \mu_{w_{2}}=\mu_{y}
$$

where $\mathrm{E}\left[z_{t}\right]=\mu_{z}, \mathrm{E}\left[w_{1, t}\right]=\mu_{w_{1}}$ and $\mathrm{E}\left[w_{2, t}\right]=\mu_{w_{2}}$, so that $\lambda_{0}+\lambda_{1} \mu_{x}=\mu_{y}$ as well from (4), and:

$$
\mathrm{E}\left[x_{t}\right]=\beta_{0}+\beta_{1} \mu_{z}+\beta_{2} \mu_{w_{2}}=\mu_{x}
$$

where from (5), $\theta_{0}+\theta_{1} \mu_{z}=\mu_{x}$. We now consider the effects of the unexpected shift in the DGP in Section 3.1.1, of the policy change in Section 3.1.2, and of mis-specification of the policy model for the DGP in Section 3.1.3.

### 3.1.1. The DGP Shifts Unexpectedly

Just as the policy change is implemented at time $T$, the DGP shifts unexpectedly to:

$$
\begin{equation*}
y_{T+1}=\gamma_{0}^{*}+\gamma_{1}^{*} x_{T+1}+\gamma_{2}^{*} w_{1, T+1}+\epsilon_{t+1} \tag{7}
\end{equation*}
$$

and:

$$
\begin{equation*}
x_{T+1}=\beta_{0}^{*}+\beta_{1}^{*} z_{T}+\beta_{2}^{*} w_{2, T+1}+v_{T+1} \tag{8}
\end{equation*}
$$

For simplicity we assume the error distributions remain the same, so that:

$$
\begin{equation*}
y_{T+1}=\left(\gamma_{0}^{*}+\gamma_{1}^{*} \beta_{0}^{*}\right)+\gamma_{1}^{*} \beta_{1}^{*} z_{T}+\gamma_{2}^{*} w_{1, T+1}+\gamma_{1}^{*} \beta_{2}^{*} w_{2, T+1}+\left(\epsilon_{T+1}+\gamma_{1}^{*} v_{T+1}\right) \tag{9}
\end{equation*}
$$

The equilibrium means after the shift and before the policy change are:

$$
\mathbf{E}_{T+1}\left[y_{T+1 \mid T}\right]=\mu_{y}^{*}=\gamma_{0}^{*}+\gamma_{1}^{*} \beta_{0}^{*}+\gamma_{1}^{*} \beta_{1}^{*} \mu_{z}+\gamma_{2}^{*} \mu_{w_{1}}+\gamma_{1}^{*} \beta_{2}^{*} \mu_{w_{2}}
$$

where the expectation is taken at time $T+1$, and when $\mu_{z}, \mu_{w_{1}}, \mu_{w_{2}}$ remain unchanged:

$$
\mu_{x}^{*}=\beta_{0}^{*}+\beta_{1}^{*} \mu_{z}+\beta_{2}^{*} \mu_{w_{2}}
$$

Even using the in-sample DGP (3) with known parameters and known values for the exogenous variables, there will be forecast failure for a large location shift from $\mu_{y}$ to $\mu_{y}^{*}$ since the forecast error $\epsilon_{T+1 \mid T}=y_{T+1}-y_{T+1 \mid T}$ is:

$$
\begin{aligned}
\epsilon_{T+1 \mid T}=\mu_{y}^{*}-\mu_{y}+ & \left(\gamma_{1}^{*} \beta_{1}^{*}-\gamma_{1} \beta_{1}\right)\left(z_{T}-\mu_{z}\right)+\left(\gamma_{2}^{*}-\gamma_{2}\right)\left(w_{1, T+1}-\mu_{w_{1}}\right) \\
& +\left(\gamma_{1}^{*} \beta_{2}^{*}-\gamma_{1} \beta_{2}\right)\left(w_{2, T+1}-\mu_{w_{2}}\right)+\left(\epsilon_{T+1}+\gamma_{1}^{*} v_{T+1}\right)
\end{aligned}
$$

so that:

$$
\begin{equation*}
\mathrm{E}_{T+1}\left[\epsilon_{T+1 \mid T}\right]=\mu_{y}^{*}-\mu_{y} \tag{10}
\end{equation*}
$$

and similarly for $v_{T+1 \mid T}=x_{T+1}-x_{T+1 \mid T}$, then $\mathrm{E}_{T+1}\left[v_{T+1 \mid T}\right]=\mu_{x}^{*}-\mu_{x}$.

### 3.1.2. The Policy Change

Next, the policy shift changes $z_{T}$ to $z_{T}+\delta$, which is a mistake as that policy leads to the impact $\delta \gamma_{1}^{*} \beta_{1}^{*}$ instead of $\delta \gamma_{1} \beta_{1}$ as anticipated by an agency using the in-sample DGP parameters. Instead, the policy shift would alter the mean forecast error to:

$$
\begin{equation*}
\mathrm{E}_{T+1}\left[\epsilon_{T+1 \mid T}\right]=\mu_{y}^{*}-\mu_{y}+\left(\gamma_{1}^{*} \beta_{1}^{*}-\gamma_{1} \beta_{1}\right) \delta \quad \text { for } \delta \neq 0 \tag{11}
\end{equation*}
$$

The additional component in (11) compared to (10) would be zero if $\delta=0$, so the failure of super exogeneity of the policy variable for the target augments, or depending on signs, possibly attenuates the forecast failure. Importantly, if $\mu_{y}^{*}=\mu_{y}$ so the DGP did not shift, the policy shift by itself could create forecast failure. Section 4 addresses testing for such a failure in advance of a policy-regime shift.

### 3.1.3. The Role of Mis-Specification

Third, there would be a mis-specification effect from using (6), as the scenario calculated ex ante would suggest the effect to be $\delta \lambda_{1} \theta_{1}$. Although there is also an estimation effect, it is probably $\mathrm{O}_{p}\left(T^{-1}\right)$, but to complete the taxonomy of forecast errors below, we will add estimation uncertainty since the parameters of (6) could not be 'known'. Denoting such estimates of the model by ${ }^{\sim}$, then:

$$
\mathrm{E}\left[y_{t}\right]=\mu_{y}=\mathrm{E}\left[\left(\widetilde{\lambda}_{0}+\widetilde{\lambda}_{1} \widetilde{\theta}_{0}\right)+\widetilde{\lambda}_{1} \widetilde{\theta}_{1} z_{t-1}\right]=\left(\lambda_{0, e}+\left(\lambda_{0} \theta_{0}\right)_{e}\right)+\left(\lambda_{1} \theta_{1}\right)_{e} \mu_{z}
$$

where the in-sample expected values of OLS estimated coefficients are shown by a subscript ${ }_{e}$, noting that both $\left\{z_{t}\right\}$ and $\mu_{z}$ are almost bound to be known to the policy agency. From the $T+1$ DGP in (9):

$$
\begin{equation*}
y_{T+1}=\mu_{y}^{*}+\gamma_{1}^{*} \beta_{1}^{*}\left(z_{T}-\mu_{z}\right)+\gamma_{2}^{*}\left(w_{1, T+1}-\mu_{w_{1}}\right)+\gamma_{1}^{*} \beta_{2}^{*}\left(w_{2, T+1}-\mu_{w_{2}}\right)+\left(\epsilon_{T+1}+\gamma_{1}^{*} v_{T+1}\right) \tag{12}
\end{equation*}
$$

but from (6), the agency would forecast $y_{T+1}$ as:

$$
\begin{equation*}
\widetilde{y}_{T+1 \mid T}=\left(\widetilde{\lambda}_{0}+\widetilde{\lambda}_{1} \widetilde{\theta}_{0}\right)+\widetilde{\lambda}_{1} \widetilde{\theta}_{1} \widetilde{z}_{T}=\widetilde{\mu}_{y}+\widetilde{\lambda}_{1} \widetilde{\theta}_{1}\left(\widetilde{z}_{T}-\mu_{z}\right) \tag{13}
\end{equation*}
$$

where $\widetilde{z}_{T}$ is the measured forecast-origin value. An agency is almost certain to know the correct value, but we allow for the possibility of an incorrect forecast-origin value to complete the taxonomy of forecast-failure outcomes in (16) and Table 1. Then the agency anticipates that shifting $z_{T}$ to $z_{T}+\delta$ would on average revise the outcome to $y_{T+1 \mid T}+\delta \widetilde{\lambda}_{1} \widetilde{\theta}_{1}$, as against the actual outcome in (12), leading to a forecast error of $\widetilde{v}_{T+1 \mid T}=y_{T+1}-\tilde{y}_{T+1 \mid T}$ :

$$
\begin{align*}
\widetilde{v}_{T+1 \mid T} & =\left(\mu_{y}^{*}-\widetilde{\mu}_{y}\right)+\left(\gamma_{1}^{*} \beta_{1}^{*}-\widetilde{\lambda}_{1} \widetilde{\theta}_{1}\right)\left(\widetilde{z}_{T}-\mu_{z}+\delta\right) \\
& +\gamma_{2}^{*}\left(w_{1, T+1}-\mu_{w_{1}}\right)+\gamma_{1}^{*} \beta_{2}^{*}\left(w_{2, T+1}-\mu_{w_{2}}\right)+\left(\epsilon_{T+1}+\gamma_{1}^{*} v_{T+1}\right) \tag{14}
\end{align*}
$$

Then (14) has an approximate average value of:

$$
\begin{equation*}
\mathrm{E}_{T+1}\left[\widetilde{v}_{T+1 \mid T}\right] \approx \mu_{y}^{*}-\mu_{y, e}+\left(\gamma_{1}^{*} \beta_{1}^{*}-\gamma_{1} \beta_{1}\right) \delta+\left(\gamma_{1} \beta_{1}-\left(\lambda_{1} \theta_{1}\right)_{e}\right) \delta \quad \text { for } \delta \neq 0 \tag{15}
\end{equation*}
$$

Unless the model is revised, the same average error will occur in the following periods leading to systematic forecast failure. Indeed, even if the policy agency included $w_{1, t+1}$ and $w_{2, t+1}$ appropriately
in their forecasting model's equations, their roles in (14) would be replaced by any shifts in their parameter values, changing (15): see Hendry and Mizon (2012).

### 3.1.4. The Sources of Forecast Failure

The error in (14) can be decomposed into terms representing mis-estimation (labelled (a)), mis-specification $((b))$ and change $((c))$ for each of the equilibrium mean (labelled (i)), slope parameter ((ii)) and unobserved terms ((iii)), as in (16). The implications of this $3 \times 3$ framework are recorded in Table 1, ignoring covariances, and under the assumption that the derived forecasts are closely similar to the direct.

$$
\begin{array}{rlr}
\widetilde{v}_{T+1 \mid T} & \simeq\left(\mu_{y, e}-\widetilde{\mu}_{y}\right) & i(a) \\
& +\left(\mu_{y}-\mu_{y, e}\right) & \\
& +\left(\mu_{y}^{*}-\mu_{y}\right) & i(c) \\
& +\left(\left(\lambda_{1} \theta_{1}\right)_{e}-\widetilde{\lambda}_{1} \widetilde{\theta}_{1}\right)\left(z_{T}-\mu_{z}+\delta\right) & i i(a) \\
& +\left(\gamma_{1} \beta_{1}-\left(\lambda_{1} \theta_{1}\right)_{e}\right)\left(z_{T}-\mu_{z}+\delta\right) & i i(b) \\
& +\left(\gamma_{1}^{*} \beta_{1}^{*}-\gamma_{1} \beta_{1}\right)\left(z_{T}-\mu_{z}+\delta\right) & i i(c) \\
& -\left(\gamma_{1}^{*} \beta_{1}^{*}-\left(\lambda_{1} \theta_{1}\right)_{e}\right)\left(z_{T}-\widetilde{z}_{T}\right) & i i i(a) \\
& +\gamma_{2}^{*}\left(w_{1, T+1}-\mu_{w_{1}}\right)+\gamma_{1}^{*} \beta_{2}^{*}\left(w_{2, T+1}-\mu_{w_{2}}\right) & i i i(b) \\
& +\left(\epsilon_{T+1}+\gamma_{1}^{*} v_{T+1}\right) & i i i(c) \tag{16}
\end{array}
$$

In (16), the impacts of each of the nine terms can be considered in isolation by setting all the others to zero in turn. Although only $i(c)$ is likely to cause forecast failure, any narrative and associated policy as discussed above may well be rejected by systematic errors arising from any of these nine terms.

### 3.2. What Can Be Learned from Systematic Forecast Failure?

The three possible mistakes-mis-estimation, mis-specification and change-potentially affect each of the three main components of any forecasting model-equilibrium means, slope parameters and unobserved terms-leading to the nine terms in Table 1. This condensed model-free taxonomy of sources of forecast errors is based on Ericsson (2017), which built on those in Clements and Hendry (1998) for closed models and Hendry and Mizon (2012) for open systems.

Table 1. A taxonomy of the implications of systematic forecast failures.

| Component | Source |  |  |
| :---: | :---: | :---: | :---: |
|  | Mis-Estimation | Mis-Specification | Change |
| Equilibrium mean | $i(a)$ | $i(b)$ | $i(c)$ |
| Reject | FN?, PV? | FM, FN, PV | FM, FN, PV |
| Slope parameter | ii(a) | $i i(b)$ | $i i(c)$ |
| Reject (if $\delta \neq 0$ ) | FN?, PV? | FM, FN, PV | FM, FN, PV |
| Unobserved terms | iii(a) [forecast origin] | iii(b) [omitted variable] | iii(c) [innovation error] |
| Reject | FN, PV | FM?, FN?, PV? | FN?, PV? |

As not all sources of systematic forecast failure lead to the same implications, we consider which sources should cause rejection of the forecasting model (denoted FM in the table), the forediction narrative (FN), or impugn policy validity (PV). Systematic failure of the forecasts could arise from any of the terms, but the consequences vary with the source of the mistake. We assume that the derived
forecasts are sufficiently close to the direct that both suffer systematic forecast failure, so in almost every case in the table, the entailed forediction and policy are rejected. However, a '?' denotes that systematic forecast failure is less likely in the given case, but if it does occur then the consequences are as shown.

Taking the columns in turn, mis-estimation by itself will rarely lead to systematic forecast failure and should not lead to rejecting the formulation of a forecasting model, or the theory from which it was derived, as appropriate estimation would avoid such problems. For example,when the forecast origin has been mis-measured by a statistical agency resulting in a large forecast error in iii(a), one should not reject either the forecasting model or the underlying theory. Nevertheless, following systematic forecast failure, any policy conclusions that had been drawn and their accompanying narrative should be rejected as incorrect, at least until more accurate data are produced.

Next, mis-specification generally entails that the model should be rejected against a less badly specified alternative, and that foredictions and policy claims must be carefully evaluated as they too will often be invalid. In stationary processes, mis-specification alone will not lead to systematic forecast failure, but the real economic world is not stationary, and policy regime shifts will, and are intended to, alter the distributions of variables. That omitted variables need not induce systematic forecast failure may surprise at first sight, but iii(b) in (16) reveals their direct effect has a zero mean as shown by Hendry and Mizon (2012).

Third, unanticipated changes in the non-zero-mean components of forecasting models will usually cause systematic forecast failure. However, slope shifts on zero-mean influences and changes in innovation error variances need not. Nevertheless, in almost all cases systematic forecast failure will induce both forediction and policy failure if the policy implemented from the mistaken narrative depended on a component that had changed. Finally, the consequences for any underlying theory depend on how that theory is formulated, an issue briefly discussed in Section 3.4. We return to the taxonomy in Section 4, where we propose a test for the elements of forediction failure.

### 3.3. A Numerical Illustration

The scenario considered is one where in (13), the agency's direct forecast, $\tilde{y}_{T+1 \mid T}$, is higher than the policy target value, and as $\widetilde{\theta}_{1}<0$, it announces a rise in $z_{T}$ of $\delta$ such that $\widetilde{y}_{T+1 \mid T}+\delta \lambda_{1} \theta_{1}$ aims to be smaller than the target, so $x_{T+1}$ will fall. Unfortunately, an unanticipated boom hits, resulting in large rises in $x_{T+1}$ and $y_{T+1}$. Below, $\widehat{y}_{T+1 \mid T}$ and $\widehat{x}_{T+1 \mid T}$ denote forecast values from (infeasible) in-sample DGP estimates, whereas $\widetilde{y}_{T+1 \mid T}$ and $\widetilde{x}_{T+1 \mid T}$ are model-based forecasts, which here also correspond to the agency's forediction-based forecasts for inflation and output. An example of how a forecast, the associated forediction and a linked policy change could be represented is given in Table 2.

Table 2. Forediction example.

| Forecast | Narrative | Policy |
| :--- | :--- | :--- |
| The outlook for inflation is | Core inflation has been elevated in <br> recent months. High levels of resource | The Policy Committee today <br> decided to raise its target |
| year followed by a moderate |  |  |
| decline to $5.25 \%$ next year. | utilization and high prices of energy <br> and commodities have the potential <br> to sustain inflationary pressures <br> but should moderate going forward. | intere rate by $\mathbf{1 0 0}$ basis points <br> due ongoing concerns <br> about inflation pressures. |

The numerical values of the parameters in the 5-variable in-sample simulation DGP are recorded in the top two rows of Table 3.

As $w_{1, t} \sim \operatorname{IN}\left[\mu_{w_{1}}, 0.1\right]$ and $w_{2, t} \sim \operatorname{IN}\left[\mu_{w_{2}}, 0.1\right]$, their projections on $x_{t}$ and $z_{t-1}$ respectively are both zero, so $\mu_{y}=6, \mu_{x}=3, \lambda_{0}=3.0, \lambda_{1}=1.0, \theta_{0}=3$ and $\theta_{1}=-1.0$. The policy change is $\delta=1$ and the shifts in the DGP are described in the bottom two rows of Table 3. Now $\mu_{y}^{*}=8.25$ and $\mu_{x}^{*}=4.5$, but initially $x$ jumps to around 6 because of the lag in its response to the policy change. For $y$
and $x$ these cases are denoted (I) \& (II) in Figure 2, whereas cases (III) \& (IV) have the same policy implementation and DGP change, but with no location shift, so $\gamma_{0}^{*}=\gamma_{0}=1.0$ and $\beta_{0}^{*}=\beta_{0}=1.0$.

Table 3. Parameter values for the simulation DGPs.

| in-sample | $\gamma_{0}$ | $\gamma_{1}$ | $\gamma_{2}$ | $\sigma_{\epsilon}^{2}$ | $\beta_{0}$ | $\beta_{1}$ | $\beta_{2}$ | $\sigma_{v}^{2}$ | $\delta$ | $\mu_{z}$ | $\mu_{w_{1}}$ | $\mu_{w_{2}}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| value | 1.0 | 1.0 | 1.0 | 0.1 | 1.0 | -1.0 | 1.0 | 0.1 | 0 | 0 | 2 | 2 |
| out-of-sample | $\gamma_{0}^{*}$ | $\gamma_{1}^{*}$ | $\gamma_{2}^{*}$ | $\sigma_{\epsilon}^{2}$ | $\beta_{0}^{*}$ | $\beta_{1}^{*}$ | $\beta_{2}^{*}$ | $\sigma_{v}^{2}$ | $\delta$ | $\mu_{z}^{*}$ | $\mu_{w_{1}}$ | $\mu_{w_{2}}$ |
| value | 2.0 | 0.5 | 2.0 | 0.1 | 2.0 | -1.5 | 2.0 | 0.1 | 1 | 1 | 2 | 2 |



Figure 2. (I,II) both forediction and policy failures from the forecast failure following a changed DGP with a location shift and a policy change combined; (III,IV) both forediction and policy failures after a changed DGP without a location shift but with a policy change.

For a single simulation, Figure 2(I) illustrates the systematic forecast failure of $\widehat{y}_{t}$ for $y_{t}$ when the break in the DGP includes both a location shift and the policy change, together with the much larger failure of the model's forecast, $\widetilde{y}_{t}$. When there is no unexpected location shift, so $\gamma_{0}=\gamma_{0}^{*}$ and $\beta_{0}=\beta_{0}^{*}$, then as (III) shows, the in-sample DGP-based forecasts $\widehat{y}_{t}$ do not suffer failure despite the other changes in the DGP. However, the model mis-specification interacting with $\delta \neq 0$ still induces both forediction and policy failure from the model's forecast $\widetilde{y}_{t}$ failing. This occurs because the agency's model for $y_{t}$ is mis-specified, so the $\delta$ change acts as an additional location shift, exacerbating the location shift in (I) from the change in $\mu_{y}$, whereas only the policy change occurs in (III). There is little difference between $\widehat{x}_{t}$ and $\widetilde{x}_{t}$ in either scenario: even if the agency had known the in-sample DGP equation for $x_{t}$, there would have been both forediction and policy failure from the changes in the slope parameters caused by $\delta \neq 0$ as Figures 2(II),(IV) show.

### 3.4. Implications of Forediction Failure for Inter-Temporal Theories

Whether or not a systematic forecast failure from whatever source impugns an underlying economic theory depends on how directly the failed model is linked to that theory. That link is close for dynamic stochastic general equilibrium (DSGE) theory, so most instances of FM in Table 1 entail a theory failure for the DSGE class. More directly, Hendry and Mizon (2014) show that the mathematical basis of inter-temporal optimization fails when there are shifts in the distributions of the variables involved. All expectations operators must then be three-way subscripted, as in $\mathrm{E}_{\mathrm{D}_{y_{t}}}\left[y_{t+1} \mid \mathcal{I}_{t-1}\right]$ which denotes the conditional expectation of the random variable $y_{t+1}$ formed at time $t$ as the integral over the distribution $\mathrm{D}_{y_{t}}(\cdot)$ given an available information set $\mathcal{I}_{t-1}$. When $\mathrm{D}_{y_{t}}(\cdot)$ is the distribution $y_{t} \sim \operatorname{IN}\left[\mu_{t}, \sigma_{y}\right]$ where future changes in $\mu_{t}$ are unpredictable, then:

$$
\begin{equation*}
\mathrm{E}_{\mathrm{D}_{y_{t}}}\left[y_{t+1} \mid \mathcal{I}_{t-1}\right]=\mu_{t} \tag{17}
\end{equation*}
$$

whereas:

$$
\begin{equation*}
\mathrm{E}_{\mathrm{D}_{y_{t+1}}}\left[y_{t+1} \mid \mathcal{I}_{t-1}\right]=\mu_{t+1} \tag{18}
\end{equation*}
$$

so the conditional expectation $\mathrm{E}_{\mathrm{D}_{y_{t}}}\left[y_{t+1} \mid \mathcal{I}_{t-1}\right]$ formed at $t$ is not an unbiased predictor of the outcome $\mu_{t+1}$, and only a 'crystal-ball' predictor $\mathrm{E}_{\mathrm{D}_{y_{t+1}}}\left[y_{t+1} \mid \mathcal{I}_{t-1}\right]$ based on 'knowing' $\mathrm{D}_{y_{t+1}}(\cdot)$ is unbiased.

A related problem afflicts derivations which incorrectly assume that the law of iterated expectations holds even though distributions shift between $t$ and $t+1$, since then only having information at time $t$ entails:

$$
\begin{equation*}
\mathrm{E}_{\mathrm{D}_{y_{t}}}\left[\mathrm{E}_{\mathrm{D}_{y_{t}}}\left[y_{t+1} \mid y_{t}\right]\right] \neq \mathrm{E}_{\mathrm{D}_{y_{t+1}}}\left[y_{t+1}\right] \tag{19}
\end{equation*}
$$

It is unsurprising that a shift in the distribution of $y_{t}$ between periods would wreck a derivation based on assuming such changes did not occur, so the collapse of DSGEs like the Bank of England's Quarterly Econometric Model (BEQEM) after the major endowment shifts occasioned by the 'Financial Crisis' should have been anticipated. A similar fate is likely to face any successor built on the same lack of mathematical foundations relevant for a changing world. Indeed, a recent analysis of the replacement model, COMPASS (Central Organising Model for Projection Analysis and Scenario Simulation) confirms that such a fate has already occurred. ${ }^{4}$

The results of this section re-emphasize the need to test forecasting and policy-analysis models for their invariance to policy changes prior to implementing them. Unanticipated location shifts will always be problematic, but avoiding the policy-induced failures seen in Figures 2(III) and (IV) by testing for invariance seems feasible. Section 4 briefly describes step-indicator saturation, then explains its extension to test invariance based on including in conditional equations the significant step indicators from marginal-model analyses.

## 4. Step-Indicator Saturation (SIS) Based Test of Invariance

Many tests for super exogeneity have been proposed since Engle et al. (1983): see e.g., (Hendry 1988; Favero and Hendry 1992; Engle and Hendry 1993; Psaradakis and Sola 1996; Jansen and Teräsvirta 1996; Hendry and Santos 2010), including tests based on co-breaking as in Krolzig and Toro (2002), and Hendry and Massmann (2007). Here we propose a test of invariance based on step-indicator saturation (denoted SIS), following Castle et al. (2015). SIS builds on the impulse-indicator saturation (IIS) approach proposed by Hendry et al. (2008) to detect outliers and shifts in a model.

[^55]In IIS, an impulse indicator is created for every observation, thereby saturating the model, but indicators are entered in feasibly large blocks for selection of significant departures. Johansen and Nielsen (2009) derive the properties of IIS for dynamic regression models, which may have unit roots, and show that parameter estimator distributions are almost unaffected despite including $T$ impulse indicators for $T$ observations, and Johansen and Nielsen (2016) derive the distribution of the gauge. Hendry and Mizon (2011) demonstrate distortions in modelling from not handling outliers, and Castle et al. (2016) show that outliers can even be distinguished from non-linear reactions as the latter hold at all observations whereas the former only occur at isolated points. Castle et al. (2015) establish that the properties of step-indicator saturation are similar to IIS, but SIS exhibits higher power than IIS when location shifts occur. Ericsson and Reisman (2012) propose combining IIS and SIS in 'super saturation' and Ericsson (2016) uses IIS to adjust an ex post mis-match between the truncated FMI Greenbook 'post-casts' and GDP outcomes. ${ }^{5}$

The invariance test proposed here is the extension to SIS of the IIS-based test of super exogeneity in Hendry and Santos (2010). SIS is applied to detect location shifts in models of policy instruments, then any resulting significant step indicators are tested in the model for the target variables of interest. Because the 'size' of a test statistic is only precisely defined for a similar test, and the word is anyway ambiguous in many settings (such as sample size), we use the term 'gauge' to denote the empirical null retention frequency of indicators. The SIS invariance test's gauge is close to its nominal significance level. We also examine its rejection frequencies when parameter invariance does not hold. When the probability of retention of relevant indicators is based on selection, it no longer corresponds to the conventional notion of 'power', so we use the term 'potency' to denote the average non-null retention frequency from selection. However, as IIS can detect failures of invariance from variance shifts in the unmodelled processes, 'super saturation' could help delineate that source of failure. Neither test requires ex ante knowledge of the timings, signs, numbers or magnitudes of location shifts, as policy-instrument models are selected automatically using Autometrics.

We test for location shifts by adding to the candidate variables a complete set of $T$ step indicators denoted $\left\{1_{\{t \leq j\}}, j=1, \ldots, T\right\}$ when the sample size is $T$, where $1_{\{t \leq j\}}=1$ for observations up to $j$, and zero otherwise (so $1_{\{t \leq T\}}$ is the intercept). Using a modified general-to-specific procedure, Castle et al. (2015) establish the gauge and the null distribution of the resulting estimator of regression parameters. A two-block process is investigated analytically, where half the indicators are added and all significant indicators recorded, then that half is dropped, and the other half examined: finally, the two retained sets of indicators are combined. The gauge, $g$, is approximately $\alpha$ when the nominal significance level of an individual test is $\alpha$. Hendry et al. (2008) and Johansen and Nielsen (2009) show that other splits, such as using $k$ splits of size $T / k$, or unequal splits, do not affect the gauge of IIS.

The invariance test then involves two stages. Denote the $n=n_{1}+n_{2}$ variables in the system by $\mathbf{q}_{t}^{\prime}=\left(\mathbf{y}_{t}^{\prime}: \mathbf{x}_{t}^{\prime}\right)$, where the $\mathbf{x}_{t}$ are the conditioning variables. Here, we only consider $n_{1}=1$. In the first stage, SIS is applied to the marginal system for all $n_{2}$ conditioning variables, and the associated significant indicators are recorded. When the intercept and $s$ lags of all $n$ variables $\mathbf{q}_{t}$ are always retained (i.e., not subject to selection), SIS is applied at significance level $\alpha_{1}$, leading to the selection of $m \geq 0$ step indicators:

$$
\begin{equation*}
\mathbf{x}_{t}=\psi_{0}+\sum_{i=1}^{s} \boldsymbol{\Psi}_{i} \mathbf{q}_{t-i}+\sum_{j=1}^{m} \eta_{j, \alpha_{1}} 1_{\left\{t \leq t_{j}\right\}}+\mathbf{v}_{2, t} \tag{20}
\end{equation*}
$$

[^56]where (20) is selected to be congruent. The coefficients of the significant step indicators are denoted $\eta_{j, \alpha_{1}}$ to emphasize their dependence on $\alpha_{1}$ used in selection. Although SIS could be applied to (20) as a system, we will focus on its application to each equation separately. The gauge of step indicators in such individual marginal models is investigated in Castle et al. (2015), who show that simulation-based distributions for SIS using Autometrics have a gauge, $g$, close to the nominal significance level $\alpha_{1}$. Section 4.1 notes their findings on the potency of SIS at the first stage to determine the occurrence of location shifts. Section 4.2 provides analytic, and Section 4.3 Monte Carlo, evidence on the gauge of the invariance test. Section 4.4 analyses a failure of invariance from a non-constant marginal process. Section 4.5 considers the second stage of the test for invariance. Section 5 investigates the gauges and potencies of the proposed automatic test in Monte Carlo experiments for a bivariate data generation process based on Section 4.4.

### 4.1. Potency of SIS at Stage 1

Potency could be judged by the selected indicators matching actual location shifts exactly or within $\pm 1, \pm 2$ periods. How often a shift at $T_{1}$ (say) is exactly matched by the correct single step indicator $1_{\left\{t \leq T_{1}\right\}}$ is important for detecting policy changes: see e.g., Hendry and Pretis (2016). However, for stage 2, finding that a shift has occurred within a few periods of its actual occurrence could still allow detection of an invariance failure in a policy model.

Analytic power calculations for a known break point exactly matched by the correct step function in a static regression show high power, and simulations confirm the extension of that result to dynamic models. A lower potency from SIS at stage 1 could be due to retained step indicators not exactly matching the shifts in the marginal model, or failing to detect one or more shifts when there are $N>1$ location shifts. The former is not very serious, as stage 1 is instrumental, so although there will be some loss of potency at stage 2 , attenuating the non-centrality parameter relative to knowing when shifts occurred, a perfect timing match is not essential for the test to reject when invariance does not hold. Failing to detect one or more shifts would be more serious as it both lowers potency relative to knowing those shifts, and removes the chance of testing such shifts at stage 2. Retention of irrelevant step indicators not corresponding to shifts in the marginal process, at a rate determined by the gauge $g$ of the selection procedure, will also lower potency but the effect of this is likely to be small for SIS.

Overall, Castle et al. (2015) show that SIS has relatively high potency for detecting location shifts in marginal processes at stage 1, albeit within a few periods either side of their starting/ending, from chance offsetting errors. Thus, we now consider the null rejection frequency at stage 2 of the invariance test for a variety of marginal processes using step indicators selected at stage 1.

### 4.2. Null Rejection Frequency of the SIS Test for Invariance at Stage 2

The $m$ significant step indicators $\left\{1_{\left\{t \leq t_{j}\right\}}\right\}$ in the equations of (20) are each retained using the criterion:

$$
\begin{equation*}
\left|\mathrm{t}_{\bar{\eta}_{j, \alpha_{1}}}\right|>c_{\alpha_{1}} \tag{21}
\end{equation*}
$$

when $c_{\alpha_{1}}$ is the critical value for significance level $\alpha_{1}$. Assuming all $m$ retained step indicators correspond to distinct shifts (or after eliminating any duplicates), for each $t$ combine them in the $m$ vector $\boldsymbol{\iota}_{t}$, and add all $\left\{\boldsymbol{\iota}_{t}\right\}, t=1, \ldots, T$ to the model for $\left\{y_{t}\right\}$ (written here with one lag):

$$
\begin{equation*}
y_{t}=\gamma_{1}+\gamma_{2}^{\prime} \mathbf{x}_{t}+\gamma_{3}^{\prime} \mathbf{q}_{t-1}+\boldsymbol{\tau}_{\alpha_{1}}^{\prime} \boldsymbol{\iota}_{t}+\boldsymbol{\epsilon}_{t} \tag{22}
\end{equation*}
$$

where $\boldsymbol{\tau}_{\alpha_{1}}^{\prime}=\left(\tau_{1, \alpha_{1}} \ldots \tau_{m, \alpha_{1}}\right)$, which should be $\mathbf{0}$ under the null, to be tested as an added-variable set in the conditional Equation (22) without selection, using an $F$-test, denoted $F_{\text {Inv }}$, at significance level $\alpha_{2}$ which rejects when $F_{\operatorname{Inv}(\tau=0)}>c_{\alpha_{2}}$. Under the null of invariance, this $F_{\operatorname{Inv}}$-test should have an approximate F -distribution, and thereby allow an appropriately sized test. Under the alternative that $\boldsymbol{\tau} \neq 0, \mathrm{~F}_{\text {Inv }}$ will have power, as discussed in Section 4.4.

Although (20) depends on the selection significance level, $\alpha_{1}$, the null rejection frequency of $F_{\operatorname{lnv}}$ should not depend on $\alpha_{1}$, although too large a value of $\alpha_{1}$ will lead to an $F$-test with large degrees of freedom, and too small $\alpha_{1}$ will lead to few, or even no, step indicators being retained from the marginal models. If no step indicators are retained in (20), so (21) does not occur, then $F_{\text {Inv }}$ cannot be computed for that data, so must under-reject in Monte Carlos when the null is true. Otherwise, the main consideration when choosing $\alpha_{1}$ is to allow power against reasonable alternatives to invariance by detecting any actual location shifts in the marginal models.

### 4.3. Monte Carlo Evidence on the Null Rejection Frequency

To check that shifts in marginal processes like (20) do not lead to spurious rejection of invariance when super exogeneity holds, the Monte Carlo experiments need to estimate the gauges, g , of the SIS-based invariance test in three states of nature for (20): (A) when there are no shifts (Section 4.3.1); (B) for a mean shift (Section 4.3.2); and (C) facing a variance change (Section 4.3.3). Values of $g$ close to the nominal significance level $\alpha_{2}$ and constant across (A)-(C) are required for a similar test. However, since Autometrics selection seeks a congruent model, insignificant irrelevant variables can sometimes be retained, and the gauge will correctly reflect that, so usually $\mathrm{g} \geq \alpha_{1}$.

The simulation DGP is the non-constant bivariate first-order vector autoregression (VAR(1)):

$$
\binom{y_{t}}{x_{t}} \left\lvert\, \mathbf{q}_{t-1} \sim \mathrm{IN}_{2}\left[\binom{\gamma_{1}+\rho \gamma_{2, t}+\kappa^{\prime} \mathbf{q}_{t-1}}{\gamma_{2, t}}, \sigma_{22}\left(\begin{array}{cc}
\sigma_{22}^{-1} \sigma_{11}+\rho^{2} \theta_{(t)} & \rho \theta_{(t)}  \tag{23}\\
\rho \theta_{(t)} & \theta_{(t)}
\end{array}\right)\right]\right.
$$

where $\mathbf{q}_{t-1}=\left(y_{t-1}: x_{t-1}\right)^{\prime}$, inducing the valid conditional relation:

$$
\begin{equation*}
\mathrm{E}\left[y_{t} \mid x_{t}, \mathbf{q}_{t-1}\right]=\gamma_{1}+\rho \gamma_{2, t}+\kappa^{\prime} \mathbf{q}_{t-1}+\frac{\rho \sigma_{22} \theta_{(t)}}{\sigma_{22} \theta_{(t)}}\left(x_{t}-\gamma_{2, t}\right)=\gamma_{1}+\rho x_{t}+\kappa^{\prime} \mathbf{q}_{t-1} \tag{24}
\end{equation*}
$$

with:

$$
\begin{equation*}
\mathrm{V}\left[y_{t} \mid x_{t}, \mathbf{q}_{t-1}\right]=\sigma_{11}+\rho^{2} \sigma_{22} \theta_{(t)}-\frac{\rho^{2} \sigma_{22}^{2} \theta_{(t)}^{2}}{\sigma_{22} \theta_{(t)}}=\sigma_{11} \tag{25}
\end{equation*}
$$

In (23), let $\gamma_{2, t}=1+\lambda 1_{\left\{t \leq T_{1}\right\}}$, so $\gamma_{2, t}$ equals $\gamma_{2}$ before $T_{1}$ and $\gamma_{2}^{*}$ after. Also $\theta_{(t)}=1+\theta 1_{\left\{t \leq T_{2}\right\}}$, so the marginal process is:

$$
\begin{equation*}
x_{t}=1+\pi 1_{\left\{t \leq T_{1}\right\}}+w_{t} \quad \text { where } w_{t} \sim \operatorname{IN}\left[0, \sigma_{22}\left(1+\theta 1_{\left\{t \leq T_{2}\right\}}\right)\right] \tag{26}
\end{equation*}
$$

As the analysis in Hendry and Johansen (2015) shows, $\mathbf{q}_{t-1}$ can be retained without selection during SIS, so we do not explicitly include dynamics as the canonical case is testing:

$$
\begin{equation*}
\mathrm{H}_{0}: \pi=0 \tag{27}
\end{equation*}
$$

in (26). Despite shifts in the marginal process, (24) and (25) show that invariance holds for the conditional model.

In the Monte Carlo, the constant and invariant parameters of interest are $\gamma_{1}=0, \rho=2, \boldsymbol{\kappa}=\mathbf{0}$ and $\sigma_{11}=1$, with $\sigma_{22}=5$, and $T_{1}=T_{2}=0.8 T$. Sample sizes of $T=(50,100,200)$ are investigated with $M=10,000$ replications, for both $\alpha_{1}$ (testing for step indicators in the marginal) and $\alpha_{2}$ (testing invariance in the conditional equation) equal to $(0.025,0.01,0.005)$, though we focus on both at 0.01 .

### 4.3.1. Constant Marginal

The simplest setting is a constant marginal process, which is (23) with $\pi=\theta=0$, so the parameters of the conditional model $y_{t} \mid x_{t}$ are $\phi_{1}^{\prime}=\left(\gamma_{1} ; \rho ; \sigma_{11}\right)=(0 ; 2 ; 1)$ and the parameters of the marginal are $\phi_{2, t}^{\prime}=\left(\gamma_{2, t} ; \sigma_{22, t}\right)=(1 ; 5)$. The conditional representation with $m$ selected indicators from (20) is:

$$
\begin{equation*}
y_{t}=\gamma_{1}+\rho x_{t}+\sum_{j=1}^{m} \tau_{j, \alpha_{1}} 1_{\left\{t \leq T_{j}\right\}}+\epsilon_{t} \tag{28}
\end{equation*}
$$

and invariance is tested by the $\mathrm{F}_{\mathrm{Inv}}$-statistic of the null $\boldsymbol{\tau}_{\alpha_{1}}=\mathbf{0}$ in (28).
Table 4 records the outcomes at $\alpha=0.01$ facing a constant marginal process, with $s=4$ lags of $(y, x)$ included in the implementation of (20). Tests for location shifts in marginal processes should not use too low a probability $\alpha_{1}$ of retaining step indicators, or else the $F_{\text {Inv }}$-statistic will have a zero null rejection frequency. For example, at $T=50$ and $\alpha_{1}=0.01$, under the null, half the time no step indicators will be retained, so only about $0.5 \alpha_{2}$ will be found overall, as simulation confirms. Simulated gauges and nominal null rejection frequencies for $F_{\text {Inv }}$ were close so long as $\alpha_{1} T \geq 2$.

Table 4. SIS simulations under the null of super exogeneity for a constant marginal process. \% no indicators' records the percentage of replications in which no step indicators are retained, so stage 2 is redundant. Stage 2 gauge records the probability of the $\mathrm{F}_{\text {Inv }}$-test falsely rejecting for the included step indicators at $\alpha_{2}=0.01$.

| $\alpha_{\mathbf{1}}=\mathbf{0 . 0 1}=\alpha_{2}$ | $\boldsymbol{T}=\mathbf{5 0}$ | $\boldsymbol{T}=\mathbf{1 0 0}$ | $\boldsymbol{T}=\mathbf{2 0 0}$ |
| :--- | :---: | :---: | :---: |
| Stage 1 gauge | 0.035 | 0.033 | 0.044 |
| \% no indicators | 0.287 | 0.098 | 0.019 |
| Stage 2 gauge | 0.006 | 0.009 | 0.009 |

### 4.3.2. Location Shifts in $\left\{x_{t}\right\}$

The second DGP is given by (23) where $\pi=2,10$ with $\gamma_{2, t}=1+\pi 1_{\left\{t \leq T_{1}\right\}}, \theta=0$ and $\kappa=0$. The results are reported in Table 5. Invariance holds irrespective of the location shift in the marginal, so these experiments check that spurious rejection is not induced thereby. Despite large changes in $\pi$, when $T>100$, Table 5 confirms that gauges are close to nominal significance levels. Importantly, the test does not spuriously reject the null, and now is only slightly undersized at $T=50$ for small shifts, as again sometimes no step indicators are retained.

Table 5. SIS simulations under the null of super exogeneity for a location shift in the marginal process. Stage 1 gauge is for retained step indicators at times with no shifts; and stage 1 potency is for when the exact indicators matching step shifts are retained, with no allowance for mis-timing.

|  |  | $\boldsymbol{\pi}=\mathbf{2}$ |  |  | $\boldsymbol{\pi}=\mathbf{1 0}$ |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| $\boldsymbol{\alpha}_{\mathbf{1}}=\mathbf{0 . 0 1}$ | $\boldsymbol{T}=\mathbf{5 0}$ | $\boldsymbol{T}=\mathbf{1 0 0}$ | $\boldsymbol{T}=\mathbf{2 0 0}$ | $\boldsymbol{T}=\mathbf{5 0}$ | $\boldsymbol{T}=\mathbf{1 0 0}$ | $\boldsymbol{T}=\mathbf{2 0 0}$ |
| Stage 1 gauge | 0.034 | 0.027 | 0.043 | 0.018 | 0.018 | 0.035 |
| Stage 1 potency | 0.191 | 0.186 | 0.205 | 0.957 | 0.962 | 0.965 |
| \% no indicators | 0.067 | 0.022 | 0.000 | 0.000 | 0.000 | 0.000 |
| Stage 2 gauge: | 0.009 | 0.010 | 0.011 | 0.010 | 0.009 | 0.010 |

### 4.3.3. Variance Shifts in $\left\{x_{t}\right\}$

The third DGP given by (23) allows the variance-covariance matrix to change while maintaining the conditions for invariance, where $\theta=2,10$ with $\pi=0, \gamma_{2, t}=1$ and $\kappa=0$. Table 6 indicates that gauges are close to nominal significance levels, and again the test does not spuriously reject the null, remaining slightly undersized at $T=50$ for small shifts.

Overall, the proposed $F_{\text {Inv }}$ test has appropriate empirical null retention frequencies for both constant and changing marginal processes, so we now turn to its ability to detect failures of invariance.

Table 6. SIS simulations under the null of super exogeneity for a variance shift in the marginal process. Legend as for Table 5.

|  | $\boldsymbol{\theta = \mathbf { 2 }}$ |  |  |  | $\boldsymbol{\theta}=\mathbf{1 0}$ |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\alpha_{\mathbf{1}}=\mathbf{0 . 0 1}$ | $\boldsymbol{T}=\mathbf{5 0}$ | $\boldsymbol{T}=\mathbf{1 0 0}$ | $\boldsymbol{T}=\mathbf{2 0 0}$ | $\boldsymbol{T}=\mathbf{5 0}$ | $\boldsymbol{T}=\mathbf{1 0 0}$ | $\boldsymbol{T}=\mathbf{2 0 0}$ |  |
| Stage 1 gauge | 0.042 | 0.051 | 0.067 | 0.060 | 0.083 | 0.113 |  |
| Stage 1 potency | 0.030 | 0.030 | 0.035 | 0.041 | 0.043 | 0.071 |  |
| \% no indicators | 0.381 | 0.135 | 0.015 | 0.342 | 0.091 | 0.005 |  |
| Stage 2 gauge | 0.006 | 0.008 | 0.009 | 0.006 | 0.009 | 0.010 |  |

### 4.4. Failure of Invariance

In this section, we derive the outcome for an invariance failure in the conditional model when the marginal process is non-constant due to a location shift, and obtain the non-centrality and approximate power of the invariance test in the conditional model for a single location shift in the marginal.

From (23) when invariance does not hold:

$$
\binom{y_{t}}{x_{t}} \left\lvert\, \mathbf{q}_{t-1} \sim \mathrm{IN}_{2}\left[\binom{\gamma_{1}+\rho \gamma_{2, t}+\kappa^{\prime} \mathbf{q}_{t-1}}{\gamma_{2, t}},\left(\begin{array}{ll}
\sigma_{11} & \sigma_{12}  \tag{29}\\
\sigma_{12} & \sigma_{22}
\end{array}\right)\right]\right.
$$

so letting $\sigma_{12} / \sigma_{22}=\beta$ leads to the conditional relation:

$$
\begin{equation*}
\mathrm{E}\left[y_{t} \mid x_{t}, \mathbf{q}_{t-1}\right]=\gamma_{1}+(\rho-\beta) \gamma_{2, t}+\beta x_{t}+\boldsymbol{\kappa}^{\prime} \mathbf{q}_{t-1} \tag{30}
\end{equation*}
$$

which depends on $\gamma_{2, t}$ when $\rho \neq \beta$ and:

$$
\begin{equation*}
x_{t}=\gamma_{2, t}+v_{2, t} \tag{31}
\end{equation*}
$$

When the dynamics and timings and forms of shifts in (31) are not known, we model $x_{t}$ using (20). Autometrics with SIS will be used to select significant regressors as well as significant step indicators from the saturating set $\sum_{j=1}^{T} \eta_{j, \alpha_{1}} 1_{\{t \leq j\}}$.

Although SIS can handle multiple location shifts, for a tractable analysis, we consider the explicit single alternative (a non-zero intercept in (32) would not alter the analysis):

$$
\begin{equation*}
\gamma_{2, t}=\pi 1_{\left\{t \leq T_{1}\right\}} \tag{32}
\end{equation*}
$$

Further, we take $\gamma_{1}$ as constant in (30), so that the location shift is derived from the failure of conditioning when $\rho \neq \beta$ :

$$
\begin{equation*}
\mathrm{E}\left[y_{t} \mid x_{t}, \mathbf{q}_{t-1}\right]=\gamma_{1}+(\rho-\beta) \pi 1_{\left\{t \leq T_{1}\right\}}+\beta x_{t}+\kappa^{\prime} \mathbf{q}_{t-1} \tag{33}
\end{equation*}
$$

The power of a test of the significance of $1_{\left\{t \leq T_{1}\right\}}$ in a scalar process when the correct shift date is known, so selection is not needed, is derived in Castle et al. (2015), who show that the test power rises with the magnitude of the location shift and the length it persists, up to half the sample. Their results apply to testing $\pi=0$ in (32) and would also apply to testing $(\rho-\beta) \pi=0$ in (33) when $1_{\left\{t \leq T_{1}\right\}}$ is known or is correctly selected at Stage 1. Section 4.5 considers the impact on the power of the invariance test of $(\rho-\beta) \pi=0$ of needing to discover the shift indicator at Stage 1. Castle et al. (2015) also examine the effects on the potency at Stage 1 of selecting a step indicator that does not precisely match the location shift in the DGP, which could alter the rejection frequency at Stage 2.

### 4.5. Second-Stage Test

The gauge of the $F_{\text {Inv }}$ test at the second stage conditional on locating the relevant step indicators corresponding exactly to the shifts in the marginal was calculated above. A relatively loose $\alpha_{1}$ will
lead to retaining some 'spurious' step indicators in the marginal, probably lowering potency slightly. In practice, there could be multiple breaks in different marginal processes at different times, which may affect one or more $x_{j, t}$, but little additional insight is gleaned compared to the one-off break in (32), since the proposed test is an F-test on all retained step indicators, so does not assume any specific shifts at either stage. The advantage of using the explicit alternative in (32) is that approximate analytic calculations are feasible. We only consider a bivariate VAR explicitly, where $(\rho-\beta)=0$ in (30) under the null that the conditional model of $\left\{y_{t}\right\}$ is invariant to the shift in $x_{t}$.

Let SIS selection applied to the marginal model for $x_{t}$ yield a set of step indicators $\mathcal{S}_{\alpha_{1}}$ defined by:

$$
\begin{equation*}
\mathcal{S}_{\alpha_{1}}=\left\{\mathrm{t}_{\eta_{i, \alpha_{1}}=0}^{2}>c_{\alpha_{1}}^{2}\right\} \tag{34}
\end{equation*}
$$

which together entail retaining $\left\{1_{\left\{t \leq t_{i}\right\}}, i=1, \ldots, m\right\}$. Combine these significant step indicators in a vector $\iota_{t}$, and add $\boldsymbol{\iota}_{t}$ to the assumed constant relation:

$$
\begin{equation*}
y_{t}=\gamma_{1}+\beta x_{t}+\boldsymbol{\kappa}^{\prime} \mathbf{q}_{t-1}+v_{1, t} \tag{35}
\end{equation*}
$$

to obtain the test regression written as:

$$
\begin{equation*}
y_{t}=\mu_{0}+\mu_{1} x_{t}+\mu_{3}^{\prime} \mathbf{q}_{t-1}+\boldsymbol{\tau}^{\prime} \iota_{t}+e_{t} \tag{36}
\end{equation*}
$$

As with IIS, a difficulty in formalizing the analysis is that the contents of $\boldsymbol{\iota}_{t}$ vary between draws, as it matters how close the particular relevant step indicators retained are to the location shift, although which irrelevant indicators are retained will not matter greatly. However, Hendry and Pretis (2016) show that the main reason SIS chooses indicators that are incorrectly timed is that the shift is either initially 'camouflaged' by opposite direction innovation errors, or same-sign large errors induce an earlier selection. In both cases, such mistakes should only have a small impact on the second-stage test potency, as simulations confirm. Failing to detect a shift in the marginal model will lower potency when that shift in fact leads to a failure of invariance in the conditional model.

### 4.6. Mis-Timed Indicator Selection in the Static Bivariate Case

We first consider the impact of mis-timing the selection of an indicator for this location shift in the conditional process in (30) derived from the non-invariant system (29) when $\boldsymbol{\kappa}=\mathbf{0}$. The conditional relation is written for $(\rho-\beta) \pi=\phi$ as:

$$
\begin{equation*}
y_{t}=\gamma_{1}+\beta x_{t}+\phi 1_{\left\{t \leq T_{1}\right\}}+\epsilon_{t} \tag{37}
\end{equation*}
$$

where $\epsilon_{t} \sim \operatorname{IN}\left[0, \sigma_{\epsilon}^{2}\right]$, which lacks invariance to changes in $\pi$ at $T_{1}$ in the marginal model (20):

$$
\begin{equation*}
x_{t}=\pi 1_{\left\{t \leq T_{1}\right\}}+v_{2, t} \tag{38}
\end{equation*}
$$

However, (37) is modelled by:

$$
\begin{equation*}
y_{t}=\mu_{0}+\mu_{1} x_{t}+\tau 1_{\left\{t \leq T_{0}\right\}}+e_{t} \tag{39}
\end{equation*}
$$

where $T_{0} \neq T_{1}$.
As SIS seeks the best matching step indicator for the location shift, any discrepancy between $1_{\left\{t \leq T_{1}\right\}}$ and $1_{\left\{t \leq T_{0}\right\}}$ is probably because the values of $\left\{\epsilon_{t}\right\}$ between $T_{0}$ and $T_{1}$ induced the mistaken choice. Setting $T_{1}=T_{0}+T_{+}$where $T_{+}>0$ for a specific formulation, then:

$$
\begin{equation*}
e_{t}=\left(\gamma_{1}-\mu_{0}\right)+\left(\left(\rho-\mu_{1}\right) \pi-\tau\right) 1_{\left\{t \leq T_{0}\right\}}+\left(\rho-\mu_{1}\right) \pi 1_{\left\{T_{0}+1 \leq t \leq T_{1}\right\}}+\epsilon_{t}+\left(\beta-\mu_{1}\right) v_{2, t} \tag{40}
\end{equation*}
$$

For observations beyond $T_{1}$ :

$$
e_{t}=\left(\gamma_{1}-\mu_{0}\right)+\left(\beta-\mu_{1}\right) v_{2, t}+\epsilon_{t}
$$

so that:

$$
\begin{equation*}
\mathrm{E}\left[\sum_{t=T_{1}+1}^{T} e_{t}^{2}\right]=\left(T-T_{1}\right)\left[\left(\gamma_{1}-\mu_{0}\right)^{2}+\left(\beta-\mu_{1}\right)^{2} \sigma_{v_{2}}^{2}+\sigma_{\epsilon}^{2}\right] \tag{41}
\end{equation*}
$$

which could be large for $\gamma_{1} \neq \mu_{0}$ and $\beta \neq \mu_{1}$. To a first approximation, least-squares selection would drive estimates to minimize the first two terms in (41). If they vanish, for $t \leq T_{0}$, (40) becomes:

$$
\begin{equation*}
e_{t}=(\phi-\tau) 1_{\left\{t \leq T_{0}\right\}}+\epsilon_{t} \tag{42}
\end{equation*}
$$

and between $T_{0}$ and $T_{1}$ :

$$
\begin{equation*}
e_{t}=\phi 1_{\left\{T_{0}+1 \leq t \leq T_{1}\right\}}+\epsilon_{t} \tag{43}
\end{equation*}
$$

suggesting SIS would find $\tau \approx \phi$, and that chance draws of $\left\{\epsilon_{t}\right\}$ must essentially offset $\phi 1_{\left\{T_{0}+1 \leq t \leq T_{1}\right\}}$ during the non-overlapping period $T_{0}+1$ to $T_{1}$. In such a setting:

$$
\mathrm{E}\left[\frac{1}{T} \sum_{t=1}^{T} e_{t}^{2}\right] \approx \sigma_{\epsilon}^{2}+\frac{\left(T_{1}-T_{0}\right)}{T} \phi^{2}
$$

so that the estimated equation error variance will not be greatly inflated by the mis-match in timing, and the rejection frequency of a t-test of $\mathrm{H}_{0}: \tau=0$ will be close to that of the corresponding test of $\mathrm{H}_{0}$ : $\phi=0$ in (37) despite the selection of the relevant indicator by SIS.

## 5. Simulating the Potencies of the SIS Invariance Test

The potency of the $\mathrm{F}_{\text {Inv }}$-test of $\mathrm{H}_{0}: \tau=0$ in (39) depends on the strength of the invariance violation, $\rho-\beta$; the magnitude of the location shift, $\pi$, both directly and through its detectability in the marginal model, which in turn depends on $\alpha_{1}$; the sample size $T$; the number of periods $T_{1}$ affected by the location shift; the number of irrelevant step indicators retained (which will affect the test's degrees of freedom, again depending on $\alpha_{1}$ ); how close the selected step shift is to matching the DGP shift; and on $\alpha_{2}$. These properties are now checked by simulation, and contrasted in experiments with the optimal, but generally infeasible, test based on adding the precisely correct indicator $1_{\left\{t \leq T_{1}\right\}}$, discussed above.

The simulation analyses used the bivariate relationship in Section 4.3 for violations of super exogeneity due to a failure of weak exogeneity under non-constancy in:

$$
\binom{y_{t}}{x_{t}} \sim \mathbb{I N}_{2}\left[\binom{\gamma_{1}+\rho \gamma_{2, t}}{\gamma_{2, t}},\left(\begin{array}{cc}
21 & 10  \tag{44}\\
10 & 5
\end{array}\right)\right]
$$

where $\gamma_{1}=2$ and $\omega^{2}=\sigma_{11}-\sigma_{12}^{2} / \sigma_{22}=1$, but $\beta=\sigma_{12} / \sigma_{22}=2 \neq \rho$, with the level shift at $T_{1}$ in the marginal $\gamma_{2, t}=\pi 1_{\left\{t>T_{1}\right\}}$ (in a policy context, it is more convenient to use $1_{\left\{t>T_{1}\right\}}=1-1_{\left\{t \leq T_{1}\right\}}$ ) so:

$$
\begin{equation*}
\gamma_{1}+\rho \gamma_{2, t}=\gamma_{1}+\rho \pi 1_{\left\{t>T_{1}\right\}} \tag{45}
\end{equation*}
$$

We vary $d=\pi / \sqrt{\sigma_{22}}$ over $1,2,2.5,3$ and $4 ; \rho$ over $0.75,1,1.5$ and 1.75 when $\beta=2$, reducing the departure from weak exogeneity; a sample size of $T=100$ with a break point at $T_{1}=80$; and significance levels $\alpha_{1}=0.01$ and $\alpha_{2}=0.01$ in the marginal and conditional, with $M=1000$ replications.

### 5.1. Optimal Infeasible Indicator-Based F-Test

The optimal infeasible F-test with a known location shift in the marginal process is computable in simulations. Table 7 reports the rejections of invariance, which are always high for large shifts, but fall as departures from weak exogeneity decrease. Empirical rejection frequencies approximate maximum achievable power for this type of test. The correct step indicator is almost always significant in the conditional model for location shifts larger than $2.5 \sqrt{\sigma_{22}}$, even for relatively small values of $(\rho-\beta)$.

Table 7. Power of the optimal infeasible F-test for a failure of invariance using a known step indicator for $\alpha_{2}=0.01$ at $T_{1}=80, T=100, M=1,000$.

| d: $\boldsymbol{\rho}$ | $\mathbf{0 . 7 5}$ | $\mathbf{1}$ | $\mathbf{1 . 5}$ | $\mathbf{1 . 7 5}$ |
| :--- | :---: | :---: | :---: | :---: |
| 1 | 1.000 | 1.000 | 0.886 | 0.270 |
| 2 | 1.000 | 1.000 | 1.000 | 0.768 |
| 2.5 | 1.000 | 1.000 | 1.000 | 0.855 |
| 3 | 1.000 | 1.000 | 1.000 | 0.879 |
| 4 | 1.000 | 1.000 | 1.000 | 0.931 |

### 5.2. Potency of the SIS-Based Test

Table 8 records the Stage 1 gauge and potency at different levels of location shift (d) and departures from weak (and hence super) exogeneity via $(\rho-\beta)$. The procedure is slightly over-gauged at Stage 1 for small shifts, when its potency is also low, and both gauge and potency are correctly unaffected by the magnitude of $(\rho-\beta)$, whereas Stage 1 potency rises rapidly with $d$.

Table 8. Stage 1 gauge and potency at $\alpha_{1}=0.01$ for $T_{1}=80, T=100, M=1000$ and $\beta=2$.

|  | Stage 1 Gauge |  |  |  | Stage 1 Potency |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\mathrm{d}: \rho$ | $\mathbf{0 . 7 5}$ | $\mathbf{1}$ | $\mathbf{1 . 5}$ | $\mathbf{1 . 7 5}$ | $\mathbf{0 . 7 5}$ | $\mathbf{1}$ | $\mathbf{1 . 5}$ | $\mathbf{1 . 7 5}$ |
| 1 | 0.038 | 0.040 | 0.041 | 0.039 | 0.231 | 0.223 | 0.227 | 0.204 |
| 2 | 0.029 | 0.028 | 0.030 | 0.030 | 0.587 | 0.575 | 0.603 | 0.590 |
| 2.5 | 0.026 | 0.026 | 0.025 | 0.025 | 0.713 | 0.737 | 0.730 | 0.708 |
| 3 | 0.023 | 0.023 | 0.024 | 0.025 | 0.820 | 0.813 | 0.803 | 0.817 |
| 4 | 0.020 | 0.021 | 0.020 | 0.022 | 0.930 | 0.930 | 0.922 | 0.929 |

Table 9 records Stage 2 potency for the three values of $\alpha_{1}$. It shows that for a failure of invariance, even when $\rho-\beta=0.25$, test potency can increase at tighter Stage 1 significance levels, probably by reducing the retention rate of irrelevant step indicators. Comparing the central panel with the matching experiments in Table 7, there is remarkably little loss of rejection frequency from selecting indicators by SIS at Stage 1, rather than knowing them, except at the smallest values of $d$.

Table 9. Stage 2 potency for a failure of invariance at $\alpha_{2}=0.01, T_{1}=80, T=100$, and $M=1000$.

|  | $\alpha_{\mathbf{1}}=\mathbf{0 . 0 2 5}$ |  |  |  | $\alpha_{\mathbf{1}}=\mathbf{0 . 0 1}$ |  |  |  | $\alpha_{\mathbf{1}}=\mathbf{0 . 0 0 5}$ |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\boldsymbol{d}: \boldsymbol{\rho}$ | $\mathbf{0 . 7 5}$ | $\mathbf{1}$ | $\mathbf{1 . 5}$ | $\mathbf{1 . 7 5}$ | $\mathbf{0 . 7 5}$ | $\mathbf{1}$ | $\mathbf{1 . 5}$ | $\mathbf{1 . 7 5}$ | $\mathbf{0 . 7 5}$ | $\mathbf{1}$ | $\mathbf{1 . 5}$ | $\mathbf{1 . 7 5}$ |
| 1 | 0.994 | 0.970 | 0.378 | 0.051 | 0.918 | 0.908 | 0.567 | 0.127 | 0.806 | 0.798 | 0.535 | 0.123 |
| 2 | 1.000 | 1.000 | 0.966 | 0.355 | 1.000 | 1.000 | 0.994 | 0.604 | 0.999 | 0.999 | 0.997 | 0.653 |
| 2.5 | 1.000 | 1.000 | 0.995 | 0.499 | 1.000 | 1.000 | 0.999 | 0.744 | 1.000 | 0.999 | 0.998 | 0.786 |
| 3 | 1.000 | 1.000 | 0.999 | 0.594 | 1.000 | 1.000 | 1.000 | 0.821 | 0.999 | 0.999 | 1.000 | 0.861 |
| 4 | 1.000 | 1.000 | 0.998 | 0.712 | 1.000 | 1.000 | 0.999 | 0.912 | 0.999 | 1.000 | 0.999 | 0.942 |

## 6. Application to the Small Artificial-Data Policy Model

To simulate a case of invariance failure from a policy change, which could be checked by $F_{\operatorname{lnv}(\boldsymbol{\tau}=0)}$ in-sample, followed by forecast failure, we splice the two scenarios from Figure 2 sequentially in the order of the 100 observations in panels (III +IV ) then those used for panels (I +II ), creating a sample of $T=200$.

Next, we estimate (5) with SIS, retaining the policy variable, and test the significance of the selected step-indicators in (6). At Stage 1 , using $\alpha_{1}=0.001$, as the model is mis-specified and the sample is $T=189$ keeping the last 11 observations for the forecast period, two indicators are selected. Testing these in $(6)$ yields $\mathrm{F}_{\operatorname{lnv}(\tau=0)}(2,186)=13.68^{* *}$, strongly rejecting invariance of the parameters of the model for $y_{t}$ to shifts in the model of $x_{t}$.

Figure 3 reports the outcome graphically, where the ellipses show the period with the earlier break in the DGP without a location shift but with the policy change. Panel (I) shows the time series for $x_{t}$ with the fitted and forecast values, denoted $\widetilde{x}_{t}$, from estimating the agency's model with SIS which delivered the indicators for testing invariance. Panel (II) shows the outcome for $y_{t}$ after adding the selected indicators denoted SIS ${ }^{\text {se }}$ from the marginal model for $x_{t}$, which had earlier rejected invariance. Panel (III) reports the outcome for a setting where the invariance failure led to an improved model, which here coincides with the in-sample DGP. This greatly reduces the later forecast errors and forediction failures. Finally, Panel (IV) augments the estimated in-sample DGP equation (with all its regressors retained) by selecting using SIS at $\alpha=0.01$. This further reduces forecast failure, although constancy can still be rejected from the unanticipated location shift. If the invariance rejection had led to the development of an improved model, better forecasts, and hopefully improved foredictions and policy decisions, would have resulted. When the policy model is not known publicly (as with MPC decisions), the agency alone can conduct these tests. However, an approximate test based on applying SIS to an adequate sequence of published forecast errors could highlight potential problems.


Figure 3. (I) Forecast failure for $x_{t}$ by $\widetilde{x}_{t}$ even with SIS; (II) Forecast failure in $\widetilde{y}_{t}$ even augmented by the SIS indicators selected from the margin model for $x_{t}$; (III) Smaller forecast failure from $\widehat{y}_{t}$ based on the in-sample DGP; (IV) Least forecast failure from $\widehat{y}_{t}$ based on the in-sample DGP with SIS.

### 6.1. Multiplicative Indicator Saturation

In the preceding example, the invariance failure was detectable by SIS because the policy change created a location shift by increasing $z_{t}$ by $\delta$. A zero-mean shift in a policy-relevant derivative, would not be detected by SIS, but could be by multiplicative indicator saturation (MIS) proposed in Ericsson (2012). MIS interacts step indicators with variables as in $d_{j, t}=z_{t} 1_{\{j \leq t\}}$, so $d_{j, t}=z_{t}$ when $j \leq t$ and is zero otherwise. Kitov and Tabor (2015) have investigated its performance in detecting changes in parameters in zero-mean settings by extensive simulations. Despite the very high dimensionality of the resulting parameter space, they find MIS has a gauge close to the nominal significance level for suitably tight $\alpha$, and has potency to detect such parameter changes. As policy failure will occur
after a policy-relevant parameter shifts, advance warning thereof would be invaluable. Even though the above illustration detected a failure of invariance, it did not necessarily entail that policy-relevant parameters had changed. We now apply MIS to the $T=200$ artificial data example in Section 6 to ascertain whether such a change could be detected, focusing on potential shifts in the coefficient of $z_{t-1}$ in (6).

Selecting at $\alpha=0.005$ as there are more then 200 candidate variables yielded:

$$
\begin{equation*}
y_{t}=-\underset{(0.27)}{1.12} z_{t-1} 1_{\{t \leq 90\}}+\underset{(0.27)}{1.17} z_{t-1} 1_{\{t \leq 100\}}+\underset{(0.045)}{6.01}-\underset{(0.2)}{0.857} z_{t-1} \tag{46}
\end{equation*}
$$

with $\widehat{\sigma}=0.60$. Thus, the in-sample shift of -1 in $\delta \lambda_{1} \theta_{1}$ is found over $t=90, \ldots, 100$, warning of a lack of invariance in the key policy parameter from the earlier policy change, although that break is barely visible in the data, as shown by the ellipse in Figure 3 (II). To understand how MIS is able to detect the parameter change, consider knowing where the shift occurred and splitting your data at that point. Then you would be startled if fitting your correctly specified model separately to the different subsamples did not deliver the appropriate estimates of their DGP parameters. Choosing the split by MIS will add variability, but the correct indicator, or one close to it, should accomplish the same task.

## 7. Forecast Error Taxonomy and Associated Tests

Table 10 relates the taxonomy in Table 1 to the sources of the forecast errors from (16) to illustrate which indicator-saturation test could be used, where the order (SIS, IIS) etc. shows their likely potency.

Table 10. The taxonomy of systematic forecast failures with associated tests.

| Component | Problem |  |  |
| :---: | :---: | :---: | :---: |
|  | Mis-Estimation | Mis-Specification | Change |
| Equilibrium mean | $i(a)$ [uncertainty] | $i(b)$ [inconsistent] | $i(c)$ [shift] |
| Source | $\left(\mu_{y, e}-\widetilde{\mu}_{y}\right)$ | $+\left(\mu_{y}-\mu_{y, e}\right)$ | $+\left(\mu_{y}^{*}-\mu_{y}\right)$ |
| Test | SIS, IIS | SIS, IIS | SIS, IIS, SIS ${ }^{\text {se }}$ |
| Slope parameter | ii(a) [uncertainty] | $i i(b)$ [inconsistent] | $i i(c)$ [break] |
| Source $(\delta \neq 0)$ | $+\left(\left(\lambda_{1} \theta_{1}\right)_{e}-\tilde{\lambda}_{1} \widetilde{\theta}_{1}\right) \times$ | $\begin{gathered} +\left(\gamma_{1} \beta_{1}-\left(\lambda_{1} \theta_{1}\right)_{e}\right) \times \\ \left(z_{T}-\mu_{z}+\delta\right) \end{gathered}$ | $\begin{aligned} & +\left(\gamma_{1}^{*} \beta_{1}^{*}-\gamma_{1} \beta_{1}\right) \times \\ & \left(z_{T}-\mu_{z}+\delta\right) \end{aligned}$ |
| Test | $\begin{gathered} \left(z_{T}-\mu_{z}+\delta\right) \\ \text { SIS } \end{gathered}$ | ( SIS | MIS, SIS, IIS, SIS ${ }^{\text {se }}$ |
| Unobserved terms | iii(a) [forecast origin] | iii(b) [omitted variable] | iii(c) [innovation error] |
| Source | $\begin{gathered} -\left(\gamma_{1}^{*} \beta_{1}^{*}-\left(\lambda_{1} \theta_{1}\right)_{e}\right) \times \\ \left(z_{T}-\widetilde{z}_{T}\right) \end{gathered}$ | $+\gamma_{2}^{*}\left(w_{1, T+1}-\mu_{w_{1}}\right)$ $+\gamma_{1}^{*} \beta_{2}^{*}\left(w_{2}\right.$ ( | ${ }_{+\epsilon_{T+1}}$ |
| Test | $\left(z_{T}-\widetilde{z}_{T}\right)$ <br> SIS, IIS | $\begin{gathered} +\gamma_{1}^{*} \bar{\beta}_{2}^{*}\left(w_{2, T+1}-\mu_{w_{2}}\right) \\ \text { IIS, SIS } \end{gathered}$ | $\begin{gathered} +\gamma_{1}^{*} \nu_{T+1} \\ \text { IIS } \end{gathered}$ |

When the source of forecast failure is the equilibrium mean or forecast origin mis-estimation, then SIS is most likely to detect the systematically signed forecast errors, whereas for other unobserved terms IIS is generally best equipped to detect these changes. When the slope parameter is the source of failure for $\delta \neq 0$, then SIS is generally best, whereas when $\delta=0$, IIS might help. In practice, policy invalidity and forediction failure are probably associated with $i(c)$ and $i i(c)$, where both SIS and IIS tests for super exogeneity are valid. In this setting, policy failure can also be triggered through ii(a) and $i i(b)$ which makes an SIS test for super exogeneity again attractive. Absent a policy intervention, then zero-mean changes result in $i i(c)$, so may best be detected using multiplicative indicator saturation.

## 8. How to Improve Future Forecasts and Foredictions

Scenarios above treated direct forecasts and those derived from foredictions as being essentially the same. When a major forecast error occurs, the agency can use a robust forecasting device such as an intercept correction (IC), or differencing the forecasts, to set them 'back on track' for the next period. Although sometimes deemed 'ad hoc', Clements and Hendry (1998) show the formal basis for their success in improving forecasts. However, the foredictions that led to the wrong policy implementation cannot be fixed so easily, even if the agency's next narrative alters its story. In our example, further increases in $\delta$ will induce greater forecast failure if the policy model is unchanged: viable policy requires invariance of the model to the policy change. Nevertheless, there is a partial 'fix' to the forecast failure and policy invalidity. If the lack of invariance is invariant, so policy shifts change the model's parameters in the same way each time as in (14), the shift associated with a past policy change can be added as an IC to a forecast based on a later policy shift. We denote such an IC by SIS ${ }_{I C}^{\text {se }}$, which has the advantage that it can be implemented before experiencing forecast failure. This is shown in Figure 4(I),(II), focusing on the last 50 periods, where the policy change coincides with the location shift at observation 191. The first panel records the forecasts for a model of $y_{t}$ which includes the SIS ${ }^{s e}$ indicator for the period of the first forecast failure, and also includes SIS ${ }_{I C}^{s e}$ as an imposed IC, from $T=$ 191. Had a location shift not also occurred, SIS ${ }_{I C}^{s e}$ would have corrected the forecast for the lack of invariance, and could have been included in the policy analysis and the associated foredictions.


Figure 4. (I) Forecasts for $y_{t}$ by $\widetilde{y}_{t}^{*}$, just using SIS $^{s e}$ for the first policy-induced shift and SIS $_{I C}^{s e}$ at observation 191; (II) Forecasts for $y_{t}$ by $\widetilde{y}_{t}^{*}$ also with SIS in-sample; (III) Forecasts from $T=192$ for $y_{t}$ by $\widetilde{y}_{i, t}$ with a 1-observation IC but without SIS; (IV) Forecasts for $y_{t}$ by $\widetilde{y}_{i, t}$ also with SIS in-sample.

Figure 4 also shows how effective a conventional IC is in the present context after the shift has occurred, using a forecast denoted by $\widetilde{y}_{i, t}$. The IC is a step indicator with a value of unity from observation $t=191$ onwards when the forecast origin is $T=192$, so one observation later, the forecast error is used to estimate the location shift. Compared to the massive forecast failure seen for the models of $y_{t}$ in Figure 3 (I) \& (II), neither of the sets of forecast errors in Figure 4 (III) \& (IV) fails a constancy
test $\left(F_{\text {Chow }}(10,188)=0.34\right.$ and $\left.F_{\text {Chow }}(10,185)=0.48\right)$. The IC alone corrects most of the forecast failure, but as (IV) shows, SIS improves the in-sample tracking by correcting the earlier location-shift induced failure and improves the accuracy of the resulting forecasts. ${ }^{6}$

In real-time forecasting, these two steps could be combined, using SIS $_{\text {IC }}^{s e}$ as the policy is implemented, followed by an IC one-period later when the location shift materialises, although a further policy change is more than likely in that event. Here, the mis-specified econometric models of the relationships between the variables are unchanged, but their forecasts are very different: successful forecasts do not imply correct models.

## 9. Conclusions

We considered two potential implications of forecast failure in a policy context, namely forediction failure and policy invalidity. Although an empirical forecasting model cannot necessarily be rejected following forecast failure, when the forecasts derived from the narratives of a policy agency are very close to the model's forecasts, as Ericsson (2016) showed was true for the FOMC minutes, then forecast failure entails forediction failure. Consequently, the associated narrative and any policy decisions based thereon also both fail. A taxonomy of the sources of forecast errors showed what could be inferred from forecast failure, and was illustrated by a small artificial-data policy model.

A test for invariance and the validity of policy analysis was proposed by selecting shifts in all marginal processes using step-indicator saturation and checking their significance in the conditional model. The test was able to detect failures of invariance when weak exogeneity failed and the marginal processes changed from a location shift. Compared to the nearest corresponding experiment in Hendry and Santos (2010), the potency of $F_{\text {Inv }}$ is considerably higher for SIS at $\alpha_{2}=0.01$ than IIS at $\alpha_{2}=0.025$ (both at $\alpha_{1}=0.025$ ) as shown in Figure 5.


Figure 5. Comparison of the potency of SIS with IIS.

A test rejection outcome by $F_{\text {Inv }}$ indicates a dependence between the conditional model parameters and those of the marginals, warning about potential mistakes from using the conditional model to predict the outcomes of policy changes that alter the marginal processes by location shifts, which is a common policy scenario. Combining these two features of forecast failure with non-invariance allows forediction failure and policy invalidity to be established when they occur. Conversely, learning that the policy model is not invariant to policy changes could lead to improved models, and we also showed a 'fix' that could help mitigate forecast failure and policy invalidity.

While all the derivations and Monte Carlo experiments here have been for 1-step forecasts from static regression equations, a single location shift and a single policy change, the general

[^57]nature of the test makes it applicable when there are multiple breaks in several marginal processes, perhaps at different times. Generalizations to dynamic equations, to conditional systems, and to other non-stationary settings, probably leading to more approximate null rejection frequencies, are the focus of our present research.

Acknowledgments: This research was supported in part by grants from the Robertson Foundation (award 9907422), Institute for New Economic Thinking (grant 20029822) and Statistics Norway (through Research Council of Norway Grant 236935). We are indebted to Jurgen A. Doornik, Neil R. Ericsson, Bent Nielsen, Ragnar Nymoen, Felix Pretis, the guest editors Rocco Mosconi and Paolo Paruolo, and two anonymous referees for helpful comments on an earlier version. It is a great pleasure to participate in a Special Issue of Econometrics in honour of Søren Johansen and Katarina Juselius who have both made major contributions to understanding the theory and practice of analysing non-stationary time series, and have been invaluable long-time co-authors. All calculations and graphs used OxMetrics Doornik (2013) and PcGive Doornik and Hendry (2013), which implements Autometrics.

Author Contributions: All authors contributed equally to this research.
Conflicts of Interest: The authors declare no conflict of interest.

## References

Akram, Q. Farooq, and Ragnar Nymoen. 2009. Model Selection for Monetary Policy Analysis: How Important is Empirical Validity? Oxford Bulletin of Economics and Statistics 71: 35-68.
Bank of England. 2015. Inflation Report, August, 2015. London: Bank of England Monetary Policy Committee.
Cartwright, Nancy. 1989. Nature's Capacities and their Measurement. Oxford: Clarendon Press.
Castle, Jennifer L., Jurgen A. Doornik, David F. Hendry, and Felix Pretis. 2015. Detecting Location Shifts During Model Selection by Step-Indicator Saturation. Econometrics 3: 240-64.
Castle, Jennifer L., Jurgen A. Doornik, and David F. Hendry. 2016. Robustness and Model Selection. Unpublished paper, Economics Department, University of Oxford, Oxford, UK.
Castle, Jennifer L., and David F. Hendry. 2011. On Not Evaluating Economic Models by Forecast Outcomes. Istanbul University Journal of the School of Business Administration 40: 1-21.
Clements, Michael P., and David F. Hendry. 1998. Forecasting Economic Time Series. Cambridge: Cambridge University Press.
Clements, Michael P., and David F. Hendry. 2005. Evaluating a Model by Forecast Performance. Oxford Bulletin of Economics and Statistics 67: 931-56.
Clements, Michael P., and J. James Reade. 2016. Forecasting and Forecast Narratives: The Bank of England Inflation Reports. Discussion paper, Economics Department, University of Reading, Reading, UK.
Doornik, Jurgen A. 2007. Econometric Model Selection With More Variables Than Observations. Working paper, Economics Department, University of Oxford, Oxford, UK.
Doornik, Jurgen A. 2009. Autometrics. In The Methodology and Practice of Econometrics. Edited by Jennifer L. Castle and Neil Shephard. Oxford: Oxford University Press, pp. 88-121.
Doornik, Jurgen A. 2013. OxMetrics: An Interface to Empirical Modelling, 7th ed. London: Timberlake Consultants Press.
Doornik, Jurgen A., and David F. Hendry. 2013. Empirical Econometric Modelling using PcGive: Volume I., 7th ed. London: Timberlake Consultants Press.
Ellison, Martin, and Thomas J. Sargent. 2012. A Defense of the FOMC. International Economic Review 53: 1047-65.
Engle, Robert F., and David F. Hendry. 1993. Testing Super Exogeneity and Invariance in Regression Models. Journal of Econometrics 56: 119-39.
Engle, Robert F., David F. Hendry, and Jean-Francois Richard. 1983. Exogeneity. Econometrica 51: 277-304.
Ericsson, Neil R. 2012. Detecting Crises, Jumps, and Changes in Regime. Working paper, Federal Reserve Board of Governors, Washington, D.C., USA.
Ericsson, Neil R. 2016. Eliciting GDP Forecasts from the FOMC's Minutes Around the Financial Crisis. International Journal of Forecasting 32: 571-83.
Ericsson, Neil R. 2017. Economic Forecasting in Theory and Practice: An Interview with David F. Hendry. International Journal of Forecasting 33: 523-42.
Ericsson, Neil R., and Erica L. Reisman. 2012. Evaluating a Global Vector Autoregression for Forecasting. International Advances in Economic Research 18: 247-58.

Favero, Carlo, and David F. Hendry. 1992. Testing the Lucas Critique: A Review. Econometric Reviews 11: 265-306. Genberg, Hans, and Andrew B. Martinez. 2014. On the Accuracy and Efficiency of IMF forecasts: A Survey and some Extensions. IEO Background Paper BP/14/04, Independent Evaluation Office of the International Monetary Fund, Washington, D.C., USA.
Hendry, David F. 1988. The Encompassing Implications of Feedback versus Feedforward Mechanisms in Econometrics. Oxford Economic Papers 40: 132-49.
Hendry, David F. 2001. How Economists Forecast. In Understanding Economic Forecasts. Edited by David F. Hendry and Neil R. Ericsson. Cambridge: MIT Press, pp. 15-41.
Hendry, David F. 2004. Causality and Exogeneity in Non-stationary Economic Time Series. In New Directions in Macromodelling. Edited by A. Welfe. Amsterdam: North Holland, pp. 21-48.
Hendry, David F. 2006. Robustifying Forecasts from Equilibrium-Correction Models. Journal of Econometrics 135: 399-426.
Hendry, David F., and Søren Johansen. 2015. Model Discovery and Trygve Haavelmo's Legacy. Econometric Theory 31: 93-114.
Hendry, David F., Søren Johansen, and Carlos Santos. 2008. Automatic Selection of Indicators in a Fully Saturated Regression. Computational Statistics 33: 317-35. Erratum, 337-39.
Hendry, David F., and Hans-Marti Krolzig. 2005. The Properties of Automatic Gets Modelling. Economic Journal 115: C32-C61.
Hendry, David F., and Michael Massmann. 2007. Co-breaking: Recent Advances and a Synopsis of the Literature. Journal of Business and Economic Statistics 25: 33-51.
Hendry, David F., and Grayham E. Mizon. 2011. Econometric Modelling of Time Series with Outlying Observations. Journal of Time Series Econometrics 3:1-26. doi:10.2202/1941-1928.1100.
Hendry, David F., and Grayham E. Mizon. 2012. Open-model Forecast-error Taxonomies. In Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis. Edited by Xiaohong Chen and Norman R. Swanson. New York: Springer, pp. 219-40.
Hendry, David F., and Grayham E. Mizon. 2014. Unpredictability in Economic Analysis, Econometric Modeling and Forecasting. Journal of Econometrics 182: 186-95.
Hendry, David F., and Felix Pretis. 2016. Quantifying the Uncertainty around Break Dates in Models using Indicator Saturation. Working paper, Economics Department, Oxford University, Oxford, UK.
Hendry, David F., and Carlos Santos. 2010. An Automatic Test of Super Exogeneity. In Volatility and Time Series Econometrics. Edited by Mark W. Watson, Tim Bollerslev and Jeffrey Russell. Oxford: Oxford University Press, pp. 164-93.
Independent Evaluation Office. 2014. IMF Forecasts: Process, Quality, and Country Perspectives. Technical report. Washington: International Monetary Fund.
Jansen, Eilev S., and Timo Teräsvirta. 1996. Testing Parameter Constancy and Super Exogeneity in Econometric Equations. Oxford Bulletin of Economics and Statistics 58: 735-63.
Johansen, Søren, and Bent Nielsen. 2009. An Analysis of the Indicator Saturation Estimator as a Robust Regression Estimator. In The Methodology and Practice of Econometrics. Edited by Jennifer L. Castle and Neil Shephard. Oxford: Oxford University Press, pp. 1-36.
Johansen, Søren, and Bent Nielsen. 2016. Asymptotic Theory of Outlier Detection Algorithms for Linear Time Series Regression Models. Scandinavian Journal of Statistics 43: 321-48.
Kitov, Oleg I., and Morten N. Tabor. 2015. Detecting Structural Changes in Linear Models: A Variable Selection Approach using Multiplicative Indicator Saturation. Unpublished paper, University of Oxford, Oxford, UK.
Krolzig, Hans-Martin, and Juan Toro. 2002. Testing for Super-exogeneity in the Presence of Common Deterministic Shifts. Annales d'Economie et de Statistique 67/68: 41-71.
Pagan, Adrian R. 2003. Report on Modelling and Forecasting at the Bank of England. Bank of England Quarterly Bulletin Spring. Available online: http:/ /www.bankofengland.co.uk/archive/Documents/historicpubs/qb/ 2003/qb030106.pdf (accessed on 5 October 2015).
Psaradakis, Zacharias, and Martin Sola. 1996. On the Power of Tests for Superexogeneity and Structural Invariance. Journal of Econometrics 72: 151-75.
Pretis, Felix, James Reade, and Genaro Sucarrat. 2016. General-to-Specific (GETS) Modelling And Indicator Saturation With The R Package Gets. Working paper, 794, Economics Department, Oxford University, Oxford, UK.

Romer, Christina D., and David H. Romer. 2008. The FOMC versus the Staff: Where can Monetary Policymakers add Value? American Economic Review 98: 230-35.
Sinclair, Tara M., Pao-Lin Tien, and Edward Gamber. 2016. Do Fed Forecast Errors Matter? CAMA Working Paper No. 47/2016, Centre for Applied Macroeconomic Analysis (CAMA), Canberra, Australia.
Siviero, Stefano, and Daniele Terlizzese. 2001. Macroeconomic Forecasting: Debunking a Few Old Wives' Tales. Discussion paper 395, Research Department, Banca d'Italia, Rome, Italy.
Spanos, Aris. 2007. Curve-Fitting, the Reliability of Inductive Inference and the Error-Statistical Approach. Philosophy of Science 74: 1046-66.
Stekler, Herman O., and Hilary Symington. 2016. Evaluating Qualitative Forecasts: The FOMC Minutes, 2006-2010. International Journal of Forecasting 32: 559-70.
Stenner, Alfred J. 1964. On Predicting our Future. Journal of Philosophy 16: 415-28.
Zhang, Kun, Jiji Zhang, and Bernhard Schölkopf. 2015. Distinguishing Cause from Effect Based on Exogeneity. Available online: http:/ /arxiv.org/abs/1504.05651 (accessed on 10 October 2015).
© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http:/ / creativecommons.org/licenses/by/4.0/).

## Article

# The Relation between Monetary Policy and the Stock Market in Europe 

Helmut Lütkepohl ${ }^{1,2, *, \dagger}$ and Aleksei Netšunajev ${ }^{3,4, \dagger}$<br>1 Department of Economics, Freie Universität Berlin, 14195 Berlin, Germany<br>2 DIW Berlin, 10117 Berlin, Germany<br>3 Department of Economics and Finance, Tallinn University of Technology, 12618 Tallinn, Estonia; aleksei.netsunajev@ttu.ee<br>4 Economics and Research Department, Bank of Estonia, 15095 Tallinn, Estonia<br>* Correspondence: hluetkepohl@diw.de<br>$\dagger$ These authors contributed equally to this work.

Received: 21 March 2018; Accepted: 31 July 2018; Published: 5 August 2018


#### Abstract

We use a cointegrated structural vector autoregressive model to investigate the relation between monetary policy in the euro area and the stock market. Since there may be an instantaneous causal relation, we consider long-run identifying restrictions for the structural shocks and also used (conditional) heteroscedasticity in the residuals for identification purposes. Heteroscedasticity is modelled by a Markov-switching mechanism. We find a plausible identification scheme for stock market and monetary policy shocks which is consistent with the second-order moment structure of the variables. The model indicates that contractionary monetary policy shocks lead to a long-lasting downturn of real stock prices.


Keywords: cointegrated vector autoregression; heteroscedasticity; Markov-switching model; monetary policy analysis

JEL Classification: C32

## 1. Introduction

The interaction of monetary policy and the stock market has been studied extensively with structural vector autoregressive (VAR) models. A central problem is the identification of the structural shocks. Nowadays, a range of different tools is available for identifying structural VARs (see Kilian and Lütkepohl 2017). Therefore, different types of identifying restrictions for monetary policy and stock market shocks have been used. For example, Bjørnland and Leitemo (2009) considered a structural VAR model for the US, where long-run and short-run restrictions were combined to identify structural shocks. Such models were also used in the context of identification by heteroscedasticity (e.g., Lütkepohl and Netšunajev 2017a, 2017b; Bertsche and Braun 2018). All these studies investigated the relation between monetary policy and the stock market in the US, but they ignored that the variables may be cointegrated. In the present study, we consider the relation between monetary policy and the stock market in Europe and explicitly account for possible cointegration of the variables involved.

European monetary policy is, of course, a central topic of empirical macroeconomics (e.g., Peersman and Smets 2001). There are also studies investigating explicitly the impact of monetary policy in Europe on the stock market. Cassola and Morana (2004) found that price-stabilizing monetary policy contributes to the stability of the stock market in the euro area. Bredin et al. (2009) performed an event study and found a negative relation between UK monetary policy and stock returns, but not between German monetary policy and stock returns. Kholodilin et al. (2009) used the approach of Rigobon and Sack (2004) and found heterogeneous relations between monetary policy and stock
prices for different sectors, but the effect of an increase in the policy rate of the European Central Bank (ECB) on an aggregate stock index was reported to be negative. m identified the ECB monetary policy shocks using an event-study approach and via heteroscedasticity following Rigobon and Sack (2004). Both identification methods yielded a negative relationship between unexpected changes in policy rates and stock returns. Alessi and Kerssenfischer (2016) estimated a structural factor model using euro area data and argued that the responses of stock returns to monetary policy are larger and quicker than in a conventional small-scale structural VAR. In a more recent study, Fausch and Sigonius (2018) used different techniques to investigate the relation between ECB monetary policy and German stock returns, including an event study, a VAR analysis-where monetary policy surprises are captured by a proxy variable-and a threshold VAR model. They found a negative relation between the ECB monetary policy and stock returns in the pre-crisis period.

In this study, we use a structural vector error correction model (VECM) identified through heteroscedasticity to investigate the relation between monetary policy and the stock market. The cointegration framework developed by Johansen and Juselius (see Johansen and Juselius 1990; Johansen 1991, 1995; Juselius 2006) opens up a convenient way to impose restrictions on the long-run effects of structural shocks in structural VAR analysis, as shown by King et al. (1991). They used the Granger-Johansen representation of a VAR model (Johansen 1995) to determine the long-run effects of their shocks, and the framework is easy to combine with identifying information obtained from the second-moment structure of the model (see Lütkepohl and Velinov 2016 or Kilian and Lütkepohl 2017, chp. 14). We use this framework in our empirical investigation.

We model the conditional heteroscedasticity in the data by a Markov-switching (MS) mechanism and find a cointegrated structural VAR model for which conventional identifying restrictions are in line with the second-moment structure of the data. The impulse responses are plausible and, in particular, production and price level go down after a contractionary monetary policy shock. Although the long-run impact of a monetary policy shock on stock prices is restricted to zero, such a shock is found to have a rather long-lasting negative impact on the stock market.

The structure of this study is as follows. In the next section, the basic structural VECM is presented and the model for the second moments is discussed in Section 3. The empirical analysis is considered in Section 4 and conclusions are presented in Section 5. The appendix provides details on the data sources.

## 2. Structural Vector Error Correction Models

The time series variables of interest are collected in the $(K \times 1)$ vector $y_{t}$. The components of $y_{t}$ may be integrated and cointegrated variables. We assume that all variables are stationary $(I(0))$ or integrated of order one $(I(1))$. Assuming a cointegration rank $r, 0 \leq r \leq K$, our point of departure is the VECM form of a VAR model,

$$
\begin{equation*}
\Delta y_{t}=v+\alpha \beta^{\prime} y_{t-1}+\Gamma_{1} \Delta y_{t-1}+\cdots+\Gamma_{p-1} \Delta y_{t-p+1}+u_{t}, \tag{1}
\end{equation*}
$$

where $\Delta$ is the differencing operator such that $\Delta y_{t}=y_{t}-y_{t-1}, v$ is a $(K \times 1)$ constant intercept term, $\alpha$ is a $(K \times r)$ loading matrix of rank $r, \beta$ is a $(K \times r)$ cointegration matrix of rank $r$, and $\Gamma_{1}, \ldots, \Gamma_{p-1}$ are $(K \times K)$ coefficient matrices (see also Johansen 1995).

The reduced-form residuals $u_{t}$ are white noise, that is, $u_{t}$ is serially uncorrelated with mean zero but may have time-varying second moments. In other words, $u_{t}$ may be heteroscedastic or conditionally heteroscedastic. The structural shocks, denoted by $\varepsilon_{t}$, are obtained from the reduced-form residuals by a linear transformation $\varepsilon_{t}=B^{-1} u_{t}$ or $B \varepsilon_{t}=u_{t}$. The $(K \times K)$ transformation matrix $B$ is assumed to be such that the structural shocks are instantaneously uncorrelated. Hence, $\mathbb{E}\left(\varepsilon_{t} \varepsilon_{t}^{\prime}\right)=\Sigma_{\varepsilon_{t}}$ is a diagonal matrix.

Substituting $B \varepsilon_{t}$ for $u_{t}$ in (1), the matrix $B$ is easily recognized as the matrix of impact effects of the structural shocks. Thus, imposing restrictions directly on the impact effects means putting restrictions
on the elements of $B$. Typically, zero restrictions are imposed on $B$, which implies that certain variables do not respond instantaneously to a shock.

The long-run effects of the shocks are easily obtained through the Granger-Johansen representation (see Johansen (1995, Theorem 4.2)) of $y_{t}$ corresponding to (1),

$$
\begin{equation*}
y_{t}=\Xi \sum_{i=1}^{t} u_{i}+\Xi^{*}(L) u_{t}+\delta_{t}+y_{0}^{*} \tag{2}
\end{equation*}
$$

where

$$
\begin{equation*}
\Xi=\beta_{\perp}\left[\alpha_{\perp}^{\prime}\left(I_{K}-\sum_{i=1}^{p-1} \Gamma_{i}\right) \beta_{\perp}\right]^{-1} \alpha_{\perp}^{\prime} \tag{3}
\end{equation*}
$$

$\Xi^{*}(L) u_{t}$ is a stationary process, $\delta_{t}$ contains deterministic terms, and $y_{0}^{*}$ represents initial conditions. In (3), $\beta_{\perp}$ and $\alpha_{\perp}$ are $(K \times(K-r))$ dimensional orthogonal complements of the ( $K \times r$ ) dimensional matrices $\beta$ and $\alpha$, respectively. If the cointegration rank $r$ is zero, the orthogonal complement matrices are replaced by $(K \times K)$ identity matrices so that the long-run effects matrix becomes

$$
\begin{equation*}
\Xi=\left(I_{K}-\sum_{i=1}^{p-1} \Gamma_{i}\right)^{-1} \tag{4}
\end{equation*}
$$

The corresponding long-run effects of the structural shocks are given by $\Xi B$. Since $\alpha$ and $\beta$ have rank $r$, their orthogonal complements have rank $K-r$, implying that $\Xi$ also has rank $K-r$, and the same holds for $\Xi B$ because $B$ is an invertible matrix of full rank $K$. For a given reduced-form matrix $\Xi$, restrictions on $\Xi B$ imply restrictions for $B$ and, hence, can help identify the structural shocks. The reduced rank of the long-run effects matrix implies that there can be at most $r$ shocks without any long-run effects, corresponding to $r$ columns of zeros of $\Xi B$. In other words, only $r$ shocks can be purely transitory. Another side effect of the reduced rank of $\Xi B$ is, however, that simply counting zero restrictions is not enough to assess identification of the structural matrix $B$, as we will see in our empirical application in Section 4.

This setup for identifying structural shocks in VAR models was proposed by King et al. (1991). Introductory treatments are given by Lütkepohl (2005, chp. 9) and Kilian and Lütkepohl (2017, chp. 10). An advantage of the structural VECM setup is that only the cointegration rank is needed, which implies the rank of the long-run effects matrix $\Xi B$. Knowing that rank, the structural shocks can be properly specified through long-run restrictions. Typically the actual cointegration relations are not needed. Thus, pretesting for specific cointegration relations and even knowing the precise order of integration of specific variables is not necessary, as long as the cointegration rank is known.

There are a number of situations of special interest. For $r=0$, the matrix of long-run effects $\Xi B$ is of full rank $K$ and, hence, cannot have zero columns. Thus, for $r=0$, all $K$ structural shocks have some long-run effects. If the cointegrating rank is zero, the VECM (1) reduces to a VAR model in first differences,

$$
\Delta y_{t}=v+\Gamma_{1} \Delta y_{t-1}+\cdots+\Gamma_{p-1} \Delta y_{t-p+1}+u_{t}
$$

for which the accumulated long-run effects on the $\Delta y_{t}$ are known to be

$$
\left(I_{K}-\sum_{i=1}^{p-1} \Gamma_{i}\right)^{-1} B
$$

The accumulated effects on the first differences are just the long-run effects on the levels $y_{t}$, of course. This case was considered by Blanchard and Quah (1989), and the estimation of the structural parameters, i.e., the $B$ matrix, is particularly easy for this case (e.g., Lütkepohl (2005, chp. 9)).

If some of the components of $y_{t}$ are $I(0)$, the long-run effects matrix $\Xi$ and, hence, also $\Xi B$, has corresponding rows of zero elements because a stationary variable is not affected permanently by a shock. Formally, that can be seen by dividing up the vector

$$
y_{t}=\left[\begin{array}{l}
y_{t}^{n} \\
y_{t}^{s}
\end{array}\right]
$$

where all components of the $\left(K_{n} \times 1\right)$ vector $y_{t}^{n}$ are $I(1)$ and all components of the $\left(K_{s} \times 1\right)$ vector $y_{t}^{s}$ are $I(0)$. In this case there exists a cointegration matrix of the form

$$
\beta=\left[\begin{array}{cc}
\beta_{(11)} & 0_{K_{n} \times K_{s}} \\
0_{K_{s} \times\left(r-K_{s}\right)} & I_{K_{s}}
\end{array}\right],
$$

where $\beta_{(11)}$ is a $\left(K_{n} \times\left(r-K_{s}\right)\right)$ matrix and $0_{M \times N}$ denotes a $(M \times N)$ zero matrix. Thus, there exists an orthogonal complement of $\beta$ such that

$$
\beta_{\perp}=\left[\begin{array}{c}
\beta_{\perp}^{(1)} \\
0_{K_{s} \times(K-r)}
\end{array}\right]
$$

Hence, the last $K_{s}$ rows of $\Xi$ are rows of zeros.
There are alternative proposals for imposing long-run restrictions for identifying structural shocks in VARs. Examples are proposals by Gonzalo and Ng (2001); Fisher et al. (2000), and Pagan and Pesaran (2008). Fisher and Huh (2014) review the literature and discuss the relations between alternative approaches.

## 3. Structural VAR Models with Changes in Volatility

If the reduced-form residuals $u_{t}$ are heteroscedastic or conditionally heteroscedastic, this property can be used for identification purposes. We use the approach of Lanne et al. (2010) and Herwartz and Lütkepohl (2014), who proposed a Markov-switching (MS) mechanism for modelling volatility changes in this context. They assumed that the distribution of the error term $u_{t}$ depends on a discrete Markov process $s_{t}$, such that

$$
\begin{equation*}
u_{t} \mid s_{t} \sim\left(0, \Sigma_{u}\left(s_{t}\right)\right) \tag{5}
\end{equation*}
$$

The Markov process $s_{t}$ has states $1, \ldots, M$ and transition probabilities

$$
p_{i j}=\operatorname{Pr}\left(s_{t}=j \mid s_{t-1}=i\right), \quad i, j=1, \ldots, M .
$$

Note that only the residual covariance matrices depend on the state $s_{t}$, whereas the VAR slope parameters are not state dependent. The model captures conditional heteroscedasticity of a quite general form.

The second-moment structure can be used for structural identification if the covariance matrices can be decomposed such that

$$
\begin{equation*}
\Sigma_{u}(1)=B B^{\prime}, \quad \Sigma_{u}(m)=B \Lambda_{m} B^{\prime}, \quad m=2, \ldots, M, \tag{6}
\end{equation*}
$$

where the $\Lambda_{m}$ are $(K \times K)$ diagonal matrices. In that case, the matrix $B$ can be used to obtain the structural shocks from the reduced-form residuals, and $B$ is identified if the following condition holds, where $\lambda_{j k}$ denotes the $k$ th diagonal element of $\Lambda_{j}$ :

$$
\begin{equation*}
\forall k, l \in\{1, \ldots, K\} \quad \exists j \in\{2, \ldots, M\} \quad \text { such that } \lambda_{j k} \neq \lambda_{j l} . \tag{7}
\end{equation*}
$$

If $M=2$, the condition means that the diagonal elements of $\Lambda_{2}$ must all be distinct. Generally, the condition requires that there is sufficient heterogeneity in the volatility changes across the shocks. If this condition is satisfied, $B$ is unique up to column sign changes and column permutations. In other words, using this $B$ matrix for computing the structural shocks from the reduced-form residuals, only the sign and positioning of the shocks remain undetermined. Note, however, that this identification approach relies on the assumption that the impact effects of the shocks remain time-invariant and only the variances change over time.

Assuming a normal conditional distribution for $u_{t} \mid s_{t}$, the model can be estimated by Gaussian maximum likelihood (ML) using the algorithm described by Herwartz and Lütkepohl (2014). It may be useful to estimate the cointegration matrix $\beta$ in a first step and keep it fixed in the subsequent optimisation of the log-likelihood with respect to the remaining parameters, including the transition probabilities. Computing the Gaussian ML estimates can be a challenge for larger models with many variables, autoregressive lags, or volatility states.

The fact that the model assigns the volatility regimes endogenously is appealing. Hence, the researcher can estimate the volatility states rather than having to know or assume them. We use the model in our empirical application, which is discussed in the next section.

## 4. Monetary Policy and the Stock Market in Europe

A five-dimensional VAR model for the euro area with variables $y_{t}=\left(q_{t}, p_{t}, c_{t}, s_{t}, r_{t}\right)^{\prime}$ is considered, where $q_{t}$ is the $\log$ of an industrial production index, $p_{t}$ denotes the $\log$ of the harmonized index of consumer prices (HICP), $c_{t}$ is a $\log$ non-energy commodity price index, $s_{t}$ is the $\log$ of the real Euro Stoxx 50 stock price index, and $r_{t}$ denotes the 3 month Euribor. The set of variables corresponds to the system used by Bjørnland and Leitemo (2009) for analysing the relation between monetary policy and the stock market in the US. We use monthly data for the period 1999M1-2014M12 and, hence, avoid the period of quantitative easing in the eurozone. Further details on the variables and data sources are given in the appendix, and the time series are plotted in Figure 1.


Figure 1. Time series used in the empirical study for sample period 1999M1-2014M12.

Given the substantial impact of the financial crisis in 2008 following the collapse of Lehman Brothers, one may wonder whether a time-invariant VAR or VECM is appropriate for the full sample period. We have therefore fitted $\operatorname{VAR}(2)$ models for $y_{t}=\left(q_{t}, p_{t}, c_{t}, s_{t}, r_{t}\right)^{\prime}$, where all or some of the coefficients are allowed to vary over time. The time-variation is governed by a two-state MS mechanism. Some models are compared in Table 1, using standard model selection criteria. The abbreviations used in Table 1 follow those proposed by Krolzig (1997), that is, MSIAH stands for a VAR model with time-varying intercept, VAR slope coefficients, and error covariance matrix; MSIH abbreviates a model where only the intercept vector and the error covariance matrix are allowed to vary; and MSH signifies a model with time-invariant intercept and VAR slope coefficients but varying error covariance matrix. In all cases, two Markov states are used, which is indicated in parentheses behind the model abbreviation. A lag order of $p=2$ is suggested by the Akaike information criterion (AIC), the Schwarz Criterion (SC), and the Hannan-Quinn criterion (HQ) (see Lütkepohl (2005, sct. 4.3)) for a time-invariant VAR process for the full sample period. Thus, in Table 1, for example, VAR(2)-MSH(2) signifies a VAR model of order 2 with 2 possible error covariance regimes and time-invariant intercept and VAR slope coefficients. It turns out that such a model is favoured by the SC and HQ model selection criteria, while AIC favours a fully flexible model. Provided the quantitative easing period is excluded from the sample, it is reasonable to believe that the monetary policy regime has not changed. With that in mind, and taking into account the preferences of the SC and HQ criteria, we use the model with time-invariant intercepts and slope coefficients and allow for a time-varying error covariance matrix only in the following. We decompose the covariance matrices such that time-invariant impact effects and time-invariant impulse responses are identified.

Table 1. Comparison of Markov-switching vector autoregressive models for $y_{t}=\left(q_{t}, p_{t}, c_{t}, s_{t}, r_{t}\right)^{\prime}$.

| Model | $\log L$ | AIC | SC | HQ |
| :---: | :---: | :---: | :---: | :---: |
| VAR(2) | 2514.761 | -4889.522 | -4662.230 | -4797.449 |
| VAR(2)-MSIAH(2) | 2668.241 | -5052.483 | -4591.405 | -4858.779 |
| VAR(2)-MSIH(2) | 2617.071 | -5050.142 | -4750.452 | -4928.766 |
| VAR(2)-MSH(2) | 2611.456 | -5048.912 | -4766.421 | -4934.479 |

[^58]Based on conventional ADF tests, all five variables are classified as $I(1)$ variables. Thus, there may be cointegration among the variables, which is worth taking into account in our structural analysis. Since the VAR(2)-MSH(2) model, and thus a model with conditional heteroscedasticity, was favoured in the previous analysis, we base tests for the cointegration rank on Johansen's (1995) cointegration rank tests robustified for conditional heteroscedasticity by generating the $p$-values with a wild bootstrap algorithm, as proposed by Cavaliere et al. (2010) and further investigated by Cavaliere et al. (2018). The results are presented in Table 2 and suggest a cointegration rank of $r=2$ if a $5 \%$ significance level is used. Thus, we consider a $\operatorname{VECM}(1)$ with one lag of the differenced variables (i.e., $p-1=1$ ) and cointegration rank $r=2$ for our structural analysis.

Table 2. Testing for the cointegration rank of a VECM(1) with intercept for $y_{t}=\left(q_{t}, p_{t}, c_{t}, s_{t}, r_{t}\right)^{\prime}$.

| $\mathbb{H}_{\mathbf{0}}$ | LR Statistic | $\boldsymbol{p}$-Value |
| :---: | :---: | :---: |
| $r=0$ | 70.570 | 0.036 |
| $r \leq 1$ | 59.559 | 0.001 |
| $r \leq 2$ | 20.539 | 0.081 |
| $r \leq 3$ | 6.227 | 0.482 |
| $r \leq 4$ | 0.248 | 0.978 |

Note: $p$-values are computed using Algorithm 1 of Cavaliere et al. (2010) with 1000 bootstrap replications.
One advantage of imposing long-run restrictions on the long-run effects matrix $\Xi B$ obtained via the Granger-Johansen representation is that we do not have to take a stand on the exact cointegration relations, we just need to know the cointegration rank of the model. Imposing only the rank restriction is also in line with the original ideas behind structural VAR modelling, which imposes as few restrictions as possible. In this context, it is perhaps worth mentioning that Bjørnland and Leitemo (2009) considered a system $y_{t}=\left(q_{t}, \Delta p_{t}, \Delta c_{t}, \Delta s_{t}, r_{t}\right)^{\prime}$ for US data which effectively assumes that there is no cointegration between the variables $p_{t}, c_{t}$ and $s_{t}$ so that these variables can be included in first differences. Although we do not want to reconsider the issue for US data here, it may be worth checking whether a setup with $\Delta p_{t}, \Delta c_{t}$, and $\Delta s_{t}$ instead of the levels variables can be used for European data as well, or whether the variables are potentially cointegrated. In Table 3, we present the results of cointegration rank tests and find clear evidence of cointegration between $p_{t}, c_{t}$, and $s_{t}$. Thus, the precise model specification used by Bjørnland and Leitemo (2009) would be difficult to defend for our European data, and we use the VECM (1) model with cointegration rank $r=2$ in the following. Another point worth emphasising is that this choice even accommodates the possibility that there are stationary variables in the model. For example, if there is a suspicion that the unit root tests have indicated unit roots just because of lack of power, this may be accommodated in our VECM setup as long as the cointegration tests indicate the cointegration rank adequately.

Table 3. Testing for the cointegration rank of a $\operatorname{VECM}(1)$ with intercept for $y_{t}=\left(p_{t}, c_{t}, s_{t}\right)^{\prime}$.

| $\mathbb{H}_{\mathbf{0}}$ | LR Statistic | $\boldsymbol{p}$-Value |
| :---: | :---: | :---: |
| $r=0$ | 53.90 | 0.001 |
| $r \leq 1$ | 0.928 | 0.999 |
| $r \leq 2$ | 0.157 | 0.984 |

Note: p-values are computed using Algorithm 1 of Cavaliere et al. (2010) with 1000 bootstrap replications. The lag order of the VECM is suggested by HQ and SC.

Based on the previous MS analysis we fit a volatility model of the type discussed in Section 3 with two states of the Markov process. The model is referred to as a VECM(1)-MSH(2) in the following. Given our small sample size, considering more volatility states is unreasonable. ${ }^{1}$ The AIC, HQ, and SC values of a $\operatorname{VECM}(1)$ model without allowing for heteroscedasticity and a VECM(1)-MSH(2) model are shown in Table 4. They clearly signal that allowing for conditional heteroscedasticity improves the model fit. The values of all three model selection criteria are substantially smaller than the corresponding values for the model without heteroscedasticity. In other words, the second-moment structure may well provide useful identifying information for the structural shocks.

[^59]Table 4. Comparison of VECMs for $y_{t}=\left(q_{t}, p_{t}, c_{t}, s_{t}, r_{t}\right)^{\prime}$.

| Model | $\log L$ | AIC | SC | HQ |
| :---: | :---: | :---: | :---: | :---: |
| VECM(1) | 2505.139 | -4888.278 | -4690.210 | -4966.639 |
| VECM(1)-MSH(2) | 2595.154 | -5034.308 | -4781.040 | -5134.507 |

Note: $L$-Gaussian likelihood, $\mathrm{AIC}=-2 \log L+2 \times$ number of free parameters, $\mathrm{SC}=-2 \log L+\log T \times$ number of free parameters, $\mathrm{HQ}=-2 \log L+2 \times$ number of free parameters $\times \log (\log T)$.

In Figure 2, the smoothed state probabilities of the VECM(1)-MSH(2) model are presented. They show that the two volatility regimes change frequently throughout the sample period so that guessing the change points reliably would be difficult for a researcher. Hence, using a model which allows for endogenously assigned volatility changes is clearly an advantage over a model where the volatility states have to be prespecified by the researcher.


Figure 2. Smoothed state probabilities of VECM(1)-MSH(2) model.

To explore the identification issue, we have to consider the diagonal elements of $\Lambda_{2}$, which represent the variances of the structural shocks in the second volatility state relative to the first state, as explained in Section 3. They are displayed in Table 5, together with estimated standard errors. Four of the five estimated relative variances of the structural shocks in the second state are smaller than 1 . Hence, most structural shocks have smaller variance in the second than in the first state. In other words, the second state captures potentially periods of lower volatility. From Figure 2, it can be seen that the first regime is associated with the period around the turn of the millennium, where the internet bubble burst, and with the period 2008/2009, where the financial crisis started. These events may have generated higher volatility in some of the structural shocks. Note, however, that at this point it is difficult to associate any of the shocks identified through heteroscedasticity with economic shocks of interest because the ordering of the shocks is arbitrary. Therefore, it is difficult to argue that specific economic shocks are more volatile in state 1.

For the identification of the shocks through heteroscedasticity, the diagonal elements of $\Lambda_{2}$ have to be distinct. Although the estimated diagonal elements are all distinct, the estimation uncertainty reflected in the standard errors is rather high and, hence, the true underlying quantities may not be distinct. This uncertainty in the estimates is not surprising given the relatively small sample size. However, some of the standard errors are quite small compared to the corresponding estimates of the relative variances, so it is reasonable to assume that at least some of the diagonal elements of $\Lambda_{2}$ are distinct. Thus, there is at least some identifying information in the second moments which may be
sufficient to discriminate between competing conventional identification schemes. We emphasise that for this to hold, it is not necessary that all the $\lambda_{2 i}$ are distinct.

Table 5. Estimated relative variances of VECM(1)-MSH(2) model for $y_{t}=\left(q_{t}, p_{t}, c_{t}, s_{t}, r_{t}\right)^{\prime}$.

| Relative Variance | Estimate | Estimated Standard Error |
| :---: | :---: | :---: |
| $\lambda_{21}$ | 1.557 | 0.587 |
| $\lambda_{22}$ | 0.765 | 0.294 |
| $\lambda_{23}$ | 0.479 | 0.186 |
| $\lambda_{24}$ | 0.024 | 0.007 |
| $\lambda_{25}$ | 0.202 | 0.075 |

We are primarily interested in a monetary policy shock and a stock market shock. Therefore, we place these shocks in the last two positions of $\varepsilon_{t}$, that is, $\varepsilon_{t}=\left(\varepsilon_{1 t}, \varepsilon_{2 t}, \varepsilon_{3 t}, \varepsilon_{t}^{s m}, \varepsilon_{t}^{m p}\right)^{\prime}$. In other words, the stock market shock, $\varepsilon_{t}^{s m}$, is the fourth shock, and the monetary policy shock, $\varepsilon_{t}^{m p}$, is last. The other components of $\varepsilon_{t}$ are left unspecified and represent other shocks to the economy.

We consider the two alternative identification schemes in Table 6. Since we have a cointegration rank $r=2$, the rank of $\Xi B$ is $K-r=3$. Thus, there can be two columns of zeros in the long-run effects matrix $\Xi B$. The first identification scheme in Table 6 assumes that both the stock market shock and the monetary policy shock are purely transitory and, hence, do not have any long-run effects on any of the variables. The two shocks are distinguished by the assumption that $\varepsilon_{t}^{\text {sm }}$ does not affect the commodity price index instantaneously, but only with some delay. Hence, there is a corresponding zero in the third row of $B$. The long-run restrictions may be justified by the notion that the effects of monetary policy and stock market shocks should be transitory. In a conventional VAR analysis, effects of these shocks on the macroeconomic variables vanish over time (see Christiano et al. 1999; Bjørnland and Leitemo 2009). The restriction on the short-run effect is needed to distinguish the two shocks, which are both neutral in the long run, and it is part of the identification schemes used by Christiano et al. (1999), Bjørnland and Leitemo (2009), and others for US data. Note also that no restrictions are imposed on the first three columns of $B$ and $\Xi B$ so that the first three shocks are identified purely by the volatility changes. Since we are not interested in them, we did not ensure that they have economic interpretations.

Table 6. Identification schemes for $y_{t}=\left(q_{t}, p_{t}, c_{t}, s_{t}, r_{t}\right)^{\prime}$.


The second identification scheme is due to Bjørnland and Leitemo (2009), who used it for US data. They were also primarily interested in the last two shocks, and arbitrarily identified the first three shocks by imposing a recursive structure on their contemporaneous impact effects, which is seen by the recursive structure of the zero entries in the first three columns of $B$. Again, the last shock is specified as monetary policy shock. It is assumed to have no long-run impact on stock prices, and this distinguishes the shock from the stock market shock. Both shocks are assumed to have no impact
effects on industrial production, the price level, and the commodity price index. This assumption reflects the belief that these variables move slowly in response to $\varepsilon_{t}^{s m}$ and $\varepsilon_{t}^{m p}$. There are no further long-run restrictions.

The two identification schemes differ not only in the way they identify the shocks of interest. While the first scheme relies primarily on long-run restrictions, the second scheme imposes most restrictions on the impact effects. Without heteroscedasticity, Scheme (1) is under-identified and the second scheme is just-identified. Thus, in the absence of heteroscedasticity, they cannot be compared with statistical tests without further assumptions. However, assuming that the shocks are already identified by the second-moment structure, the zero restrictions on $B$ and $\Xi B$ are over-identifying and can be tested by standard likelihood ratio (LR) tests.

The results of such tests are presented in Table 7, where the alternative is a model that is purely identified by heteroscedasticity and has no zero restrictions on $B$ or $\Xi B$. In addition to the value of the LR statistic, the assumed degrees of freedom of the $\chi^{2}$ limiting distributions are presented on which the $p$-values are based. For both tests, the number of degrees of freedom is determined under the assumption that the structural matrix $B$ is fully identified by heteroscedasticity. The number of degrees of freedom for testing Scheme (1) is 7 because, for $r=2$, the long-run effects matrix $\Xi B$ has rank $K-r=3$ and, hence, each column of zeros counts for three independent restrictions only. In other words, the columns of the matrix $\Xi B$ can be represented as a linear combination of three basis vectors. A zero column is obtained if the basis vectors are multiplied by zero coefficients. Thus, each of the two columns of zeros is obtained by restricting the three weights of the basis vectors to zero. Hence, the 10 zero coefficients in the last two columns of the long-run effects matrix in Scheme (1) count for six restrictions only. In addition, there is one zero restriction on the impact effects matrix $B$.

Table 7. Tests of identification schemes.

| $\mathbb{H}_{\mathbf{0}}$ | LR Statistic | Degrees of Freedom | $\boldsymbol{p}$-Value |
| :---: | :---: | :---: | :---: |
| Scheme (1) | 62.494 | 7 | $4.787 \times 10^{-11}$ |
| Scheme (2) | 18.807 | 10 | 0.043 |

For Scheme (1), the $p$-value of the LR test is less than $1 \%$, so Scheme (1) is rejected at any conventional significance level. This also indicates that there must be identifying information in the second-moment structure because the $\mathbb{H}_{0}$ model is not identified by the zero restrictions alone. Thus, the heteroscedastic structure has identifying power. However, since the evidence of all relative variances being distinct in Table 5 is weak, there is of course the possibility that the structure is only partially identified by heteroscedasticity. In that case, the degrees of freedom of the LR tests in Table 7 may be smaller than assumed in the table, implying that the $p$-values would be even smaller than those in the table.

Considering the second $p$-value in Table 7, Identification Scheme (2) cannot be rejected at a $1 \%$ level, although its $p$-value is slightly below $5 \%$, at least if a $\chi^{2}$ distribution with 10 degrees of freedom is used as reference. This outcome is interesting because, in a related study based on identification through heteroscedasticity, Lütkepohl and Netšunajev (2017a) found strong evidence against the restrictions for the US. Admittedly, this evidence is based on a quite different sample period. Moreover, Bertsche and Braun (2018) did not confirm this result with a different volatility model. However, Lütkepohl and Netšunajev (2017b) also found an implausible reaction of the inflation rate to a monetary policy shock for the US. Thus, it is instructive to see the responses of the variables to the two shocks of interest for our European model when Identification Scheme (2) is used.

The estimated impulse responses with $\pm$ one standard error bootstrap confidence intervals are shown in Figure 3. The confidence intervals have a $68 \%$ level in a Gaussian environment. The standard errors are estimated with a fixed-design wild bootstrap conditioning on the transition probabilities, as proposed in Herwartz and Lütkepohl (2014) and also used in Lütkepohl and Netšunajev (2017a).


Figure 3. Responses to stock market and monetary policy shocks with $\pm$ one standard error confidence intervals for Identification Scheme (2).

The responses of the variables to both shocks are quite plausible. A stock market shock is followed by increases in all other variables, although these upswings occur with some delay. This is in line with the studies by Bjørnland and Leitemo (2009) and Li et al. (2010) for the US. The increase in output and inflation is consistent with the view that a rise in stock prices increases consumption (Beaudry and Portier 2006) through a wealth effect and investment through a Tobin Q effect, thus inducing aggregate demand to increase. Due to nominal rigidities, prices react slowly, and inflation as well as commodity prices rise in the intermediate run. The response of the interest rate may be explained by the behaviour of an inflation-targeting central bank which is increasing interest rates to combat the inflationary pressure of a high aggregate demand.

A contractionary monetary policy shock, induced by an increase in the interest rate, reduces industrial production, the price level, and commodity prices with some delay. Similar to Peersman and Smets (2001) and Ehrmann et al. (2003), we do not find any price puzzle and observe long-run effects on prices. The shock leads to a long-lasting downturn of the stock index after a monetary policy tightening. This is in line with results of Bjørnland and Leitemo (2009) for the US, but the effect for the euro area is not as pronounced as in their study. Even though there is mixed evidence regarding the influence of monetary policy on industry- or country-level stock returns in Europe, most of the studies agree on a negative relation between ECB monetary policy and aggregate stock returns in the euro area (Kholodilin et al. 2009; Alessi and Kerssenfischer 2016; Fausch and Sigonius 2018).

Clearly, our results support those previous findings which show that the policy of the ECB has a substantial impact on the stock market in Europe.

## 5. Conclusions

We have constructed a five-dimensional cointegrated structural VAR model for monthly euro area variables for the period 1999M1-2014M12 to study the relation between monetary policy and the stock market. The period of quantitative easing is explicitly excluded because it may be regarded as a new monetary policy regime. We allowed for conditional heteroscedasticity in the data and modelled volatility changes by a Markov-switching mechanism. Heteroscedasticity was used to disentangle a stock market and a monetary policy shock. Conventional identification restrictions on the impact and long-run effects of the structural shocks that have been used for a similar model for the US are found to be roughly consistent with the second-moment structure of the variables for our sample period, while an alternative identification scheme is strongly rejected.

The impulse responses for the maintained identification scheme are economically plausible and, in particular, production and price level go down after a contractionary monetary policy shock. Although the long-run impact of a monetary policy shock on stock prices is restricted to be neutral, such a shock is found to have a rather long-lasting negative impact on the stock market.

Author Contributions: The authors contributed equally to this research.
Funding: Financial support from the Estonian Research Council grant MOBTP11 is gratefully acknowledged.
Acknowledgments: We thank the guest editors and two anonymous referees for valuable comments which have improved the exposition of our paper.
Conflicts of Interest: The authors declare no conflict of interest. The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

## Appendix A. Variables and Data Sources

The industrial production index is seasonally adjusted and obtained from the ECB Statistical Data Warehouse. The harmonized index of consumer prices (HICP) is also seasonally adjusted and obtained from Eurostat. The non-energy commodity price index is obtained from the World Bank. The Euro Stoxx 50 stock price series is obtained from www.stoxx.com and deflated by the consumer price index to measure real stock prices. Finally, the 3 month Euribor is obtained from the ECB Statistical Data Warehouse.

## References

Alessi, Lucia, and Mark Kerssenfischer. 2016. The Response of Asset Prices to Monetary Policy Shocks: Stronger Than Thought. Working Paper Series 1967. Frankfurt am Main, Germany: European Central Bank.
Beaudry, Paul, and Franck Portier. 2006. Stock prices, news, and economic fluctuations. American Economic Review 96: 1293-307. [CrossRef]
Bertsche, Dominik, and Robin Braun. 2018. Identification of Structural Vector Autoregressions by Stochastic Volatility. Technical Report 2018-01. Konstanz: University of Konstanz.
Bjørnland, Hilde C., and Kai Leitemo. 2009. Identifying the interdependence between US monetary policy and the stock market. Journal of Monetary Economics 56: 275-82. [CrossRef]
Blanchard, Olivier J., and Danny Quah. 1989. The dynamic effects of aggregate demand and supply disturbances. American Economic Review 79: 655-73.
Bredin, Don, Stuart Hyde, Dirk Nitzsche, and Gerard O'Reilly. 2009. European monetary policy surprises: The aggregate and sectoral stock market response. International Journal of Finance \& Economics 14: 156-71. [CrossRef]
Cassola, Nuno, and Claudio Morana. 2004. Monetary policy and the stock market in the euro area. Journal of Policy Modeling 26: 387-99. [CrossRef]

Cavaliere, Giuseppe, Anders Rahbek, and A. M. Robert Taylor. 2010. Cointegration rank testing under conditional heteroskedasticity. Econometric Theory 26: 1719-60. [CrossRef]
Cavaliere, Giuseppe, Luca De Angelis, Anders Rahbek, and AM Robert Taylor. 2018. Determining the cointegration rank in heteroskedastic VAR models of unknown order. Econometric Theory 34: 349-82. [CrossRef]
Christiano, Lawrence J., Martin Eichenbaum, and Charles L. Evans 1999. Monetary policy shocks: What have we learned and to what end? In Handbook of Macroeconomics. Amsterdam: Elsevier, chp. 2, vol. 1, pp. 65-148. [CrossRef]
Ehrmann, Michael, Leonardo Gambacorta, Jorge Martinez-Pagés, Patrick Sevestre, and Andreas Worms. 2003. The effects of monetary policy in the euro area. Oxford Review of Economic Policy 19: 58-72. [CrossRef]
Fausch, Jürg, and Markus Sigonius. 2018. The impact of ECB monetary policy surprises on the German stock market. Journal of Macroeconomics 55: 46-63. doi;10.1016/j.jmacro.2017.09.001. [CrossRef]
Fisher, Lance A., and Hyeon-seung Huh. 2014. Identification methods in vector-error correction models: Equivalence results. Journal of Economic Surveys 28: 1-16. [CrossRef]
Fisher, Lance A., Hyeon-Seung Huh, and Peter M. Summers. 2000. Structural identification of permanent shocks in VEC models: A generalization. Journal of Macroeconomics 22: 53-68. [CrossRef]
Gonzalo, Jesus, and Serena Ng. 2001. A systematic framework for analyzing the dynamic effects of permanent and transitory shocks. Journal of Economic Dynamics and Control 25: 1527-46. [CrossRef]
Haitsma, Reinder, Deren Unalmis, and Jakob de Haan. 2016. The impact of the ECB's conventional and unconventional monetary policies on stock markets. Journal of Macroeconomics 48: 101-16. [CrossRef]
Herwartz, Helmut, and Helmut Lütkepohl. 2014. Structural vector autoregressions with Markov switching: Combining conventional with statistical identification of shocks. Journal of Econometrics 183: 104-16. [CrossRef]
Johansen, Søren, and Katarina Juselius. 1990. Maximum likelihood estimation and inference on cointegration-With applications to the demand for money. Oxford Bulletin of Economics and Statistics 52: 169-210.
Johansen, Søren. 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica 59: 1551-81. [CrossRef]
Johansen, Søren. 1995. Likelihood-based Inference in Cointegrated Vector Autoregressive Models. Oxford: Oxford University Press. [CrossRef]
Juselius, Katarina. 2006. The Cointegrated VAR Model: Methodology and Applications. Oxford: Oxford University Press.
Kholodilin, Konstantin, Alberto Montagnoli, Oreste Napolitano, and Boriss Siliverstovs. 2009. Assessing the impact of the ECB's monetary policy on the stock markets: A sectoral view. Economics Letters 105: 211-13. [CrossRef]
Kilian, Lutz, and Helmut Lütkepohl. 2017. Structural Vector Autoregressive Analysis. Cambridge: Cambridge University Press.
King, Robert, Charles I. Plosser, James H. Stock, and Mark W. Watson. 1991. Stochastic trends and economic fluctuations. American Economic Review 81: 819-40.
Krolzig, Hans-Martin. 1997. Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis. Berlin: Springer.
Lütkepohl, Helmut, and Aleksei Netšunajev. 2017a. Structural vector autoregressions with heteroskedasticy: A review of different volatility models. Econometrics and Statistics 1: 2-18. [CrossRef]
Lütkepohl, Helmut, and Aleksei Netšunajev. 2017b. Structural vector autoregressions with smooth transition in variances. Journal of Economic Dynamics and Control 84: 43-57. [CrossRef]
Lütkepohl, Helmut, and Anton Velinov. 2016. Structural vector autoregressions: Checking identifying long-run restrictions via heteroskedasticity. Journal of Economic Surveys 30: 377-92.
Lütkepohl, Helmut. 2005. New Introduction to Multiple Time Series Analysis. Berlin: Springer. [CrossRef]
Lanne, Markku, Helmut Lütkepohl, and Katarzyna Maciejowska. 2010. Structural vector autoregressions with Markov switching. Journal of Economic Dynamics and Control 34: 121-31. [CrossRef]
Li, Yun Daisy, Talan B. Iscan, and Kuan Xu. The impact of monetary policy shocks on stock prices: Evidence from Canada and the United States. Journal of International Money and Finance 29: 876-96. [CrossRef]
Pagan, Adrian R., and M. Hashem Pesaran. 2008. Econometric analysis of structural systems with permanent and transitory shocks. Journal of Economic Dynamics and Control 32: 3376-95. [CrossRef]
Peersman, Gert, and Frank Smets. 2001. The Monetary Transmission Mechanism in the Euro Area: More Evidence from VAR Analysis. Working Paper Series 0091. Frankfurt am Main, Germany: European Central Bank.

Rigobon, Roberto, and Brian Sack. 2004. The impact of monetary policy on asset prices. Journal of Monetary Economics 51: 1553-75. [CrossRef]
© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http:/ / creativecommons.org/licenses/by/4.0/).

## Article

# Cointegration and Structure in Norwegian Wage-Price Dynamics ${ }^{\dagger}$ 

Marit Gjelsvik ${ }^{1}$, Ragnar Nymoen ${ }^{2, *}$ and Victoria Sparrman ${ }^{1}$<br>1 Statistics Norway, Research Department, Postboks 2633 St.Hanshaugen, NO-0131 Oslo, Norway; Marit.Gjelsvik@hkinorge.no (M.G.); victoria.sparrman@nho.no (V.S.)<br>2 Department of Economics, University of Oslo, Postboks 1095 Blindern, NO-0317 Oslo, Norway<br>* Correspondence: ragnar.nymoen@econ.uio.no; Tel.: +47-228-551-48<br>$\dagger$ We would like to thank Ådne Cappelen, Jennifer L. Castle, David F. Hendry, Erling Holmøy, Håvard Hungnes and Anders Rygh Swensen for comments and discussions, as well as seminar participants at conferences and workshops at the University of Copenhagen, the Halle Institute for Economic Research, Norges Bank, and Statistics Norway. We wish to express thanks to the guest-editors Rocco Mosconi and Paolo Parulo, and to the two anonymous referees for comments both critical and constructive. The numerical results in the paper were obtained using PcGive 15.0 and Ox 8, provided by Oxmetrics, Oxford, UK.

Received: 24 March 2018; Accepted: 8 July 2020; Published: 27 July 2020


#### Abstract

Wage coordination plays an important role in macroeconomic stabilization. Pattern wage bargaining systems have been common in Europe, but in different forms, and with different degrees of success in terms of actual coordination reached. We focus on wage formation in Norway, a small open economy, where it is custom to regard the manufacturing industry as the wage leader. We estimate a model of wage formation in manufacturing and in two other sectors. Deciding cointegration rank is an important step in the analysis, economically as well statistically. In combination with simultaneous equation modelling, the cointegration analysis provides evidence that collective wage negotiations in manufacturing have defined wage norms for the rest of the economy over the period 1980(1)-2014(4).


Keywords: cointegration; error-correcting adjustment; estimation and hypothesis testing in cointegrated models; rent-sharing in wage formation; pattern wage bargaining; inflation targeting; small open economy wage policies; inflation targeting; macroeconomic fluctuations and transmission mechanisms

JEL Classification: C52; E24; E31; E37; J31

## 1. Introduction

We investigate econometrically whether collective agreements in the manufacturing sector have defined norms for wage adjustments in the other sectors of the Norwegian economy over the period 1980(1)-2014(4). By linking the evolution of the wage cost level to economic fundamentals (e.g., business sector profitability and productivity), wage norms can give a sustainable national wage path. Wage pattern bargaining can also stabilize wage relativities at the industry level. The econometric model include both the short-run and long-term relationships of the national wage system. We therefore use the integrated and cointegrated VAR (CVAR) framework to which Johansen and Juselius have contributed significantly, see Johansen and Juselius (1990); Johansen (1995b); Juselius (2007).

There are several milestones in the history of economic research that give econometric treatment to wage formation, see Forder (2014, chp. 1-4) and Bårdsen et al. (2005, chp. 3-4). A new epoch in the econometric modelling of wages commenced with the development of the theory of cointegration, Engle and Granger (1987); Johansen (1988). "Error-correction" mechanisms in dynamic models of wage and price setting, which had been used by Denis Sargan in several important applied papers that
accompanied his seminal methodological contribution (e.g., Sargan (1964); Sargan (1980)), could now be recognized as having a clear foundation in cointegration methodology. Economic models could also be formulated that incorporated both deterministic drift and stochastic trends in wage formation, Nymoen (1989a). The new econometrics of wage modelling also clarified how restrictive the Phillips curve models had been by imposing untested unit root restrictions that forced a particular natural rate of unemployment dynamics on the system of wage and price setting, see Bårdsen and Nymoen (2003); Kolsrud and Nymoen (2014) and Kolsrud and Nymoen (2015).

One advantage of the VAR (CVAR) approach is that we can test hypotheses of absence of cointegration between wages in different industries without assuming anything about the bargaining pattern from the outset. Conditional on cointegration, we can develop empirical models of wage adjustments relevant for testing specific bargaining patterns. For example, we can investigate whether a stochastic trend in the manufacturing wage level is transmitted to the private service sector, which we would expect if the two sectors act as wage leader and wage follower.

CVARs have previously been used to estimate wage leadership models (Lindquist and Vilhelmsson (2006) (Sweden); Lamo et al. (2012); Camarero et al. (2014) (multiple countries)). Unlike these studies, we include the foreign sector, which is important for representing an imported nominal stochastic trend in the econometric model, and to test whether the importance of this trend may have changed. We also show that the simultaneous equations model (SEM) can be useful for modelling the structure of wage pattern bargaining. Nymoen (1991) specified a SEM for Norwegian wages and prices with two sectors, which was estimated by full information maximum likelihood (FIML), but without any formal analysis of cointegration rank.

Historically, the wage leader model was one of several initiatives to curb inflation in the post-war period, in a situation with full employment and with a commitment to free collective bargaining, Aukrust (1977); Meade (1982); Forder (2014) among others. Similar systems were developed in Sweden Edgren et al. (1969); France Courbis (1974) and the Netherlands Driehuis and de Wolff (1976). A key point was that in the export and import competing sectors of the economy, considerations about the required return to capital served as an automatic stabilizer of nominal wage cost growth. Over time, it was one of the corrective mechanisms that would make the wage cost level fluctuate around a growth path determined by product price and average labour productivity, which defined the scope for sustainable compensation to wage earners. The sheltered sectors, without foreign competition, could compensate for increased wages (increased production costs) by adjusting prices. In Norway, the manufacturing sector was given the role of wage leader, and if the followers were loyal to the system, they would, on average, get the value of a much higher productivity growth than if they broke out of the system. The manufacturing sector also signs the leading collective agreements in several other European countries, see Knell and Stiglbauer (2009).

However, in a recent theoretical contribution, Calmfors and Seim (2013) challenge the conventional wisdom that such pattern bargaining produces wage restraint. They show theoretically that wage restraint depends on the monetary policy regime and the size of the leading sector. This serves as a reminder that wage bargaining has a clear institutional dimension, and that institutions change over time, cf. Soskice (1990); Camarero et al. (2016). The possibility of a connection between monetary policy regimes and the system of wage and price setting has also been analysed by Cukierman and Lippi (1999); Iversen (1999); Soskice and Iversen (2000) and Holden (2005), among others.

The system of pattern wage bargaining represents an advanced product of civilization. Disruption of such institutions can occur due to changes elsewhere in the economy, or in the wider society. Hence we look for signs of structural breaks. Two important events stand out in the period covered by our data set: a change in monetary policy, and a historic increase in labour immigration. One-third of the sample used represents the era after inflation targeting was introduced in 2001, and about one quarter of the sample is from the period of high labour immigration due to EU enlargement. Hence, if the monetary policy regime change or the new immigration flow, or both, have affected wage formation with any force, the structural break should be detectable empirically.

The paper is organized as follows: In Section 2, the main theoretical model equations are specified, together with the hypotheses that we aim to test empirically. In Section 3, the data set is presented and the VAR that we use for the econometric analysis is formulated. The common ground of the pattern wage bargaining tests is provided by the cointegration analysis, and is presented in Section 4. Our estimation results for the pattern wage bargaining parameters are presented in Section 5. Our results indicate that the Norwegian wage model has preserved its main leader-follower relationships also in the period of inflation targeting, while the rise in immigration appears to have reduced the wage target that the wage setters tends to compromise. Our conclusions are also summarized in Section 6, together with a short discussion.

## 2. Wage Pattern Relationship

In practice, national wage setting takes place in mixed system, made up of collective agreements, individual contracts and where legislative measures (e.g., minimum wages) have a role to play. The Norwegian system is no exception, but is it also custom to regard it as relatively more regulated by voluntary collective agreements than is seen in most other industrial economies, Evju (2014) and Nymoen (2017).

The results of collective bargaining typically represent compromises about annual wage adjustments that balance the concerns about required profitability with fairness in the workers' share of the industry's added value. A wage path that has a common trend with the value of average labour productivity, often referred to as wage scope (or ability to pay), is then implied. Another term for this kind of relationship is rent-sharing in wage formation.

In the following, we focus on three sectors: manufacturing (labelled sector 1), the private service industry (sector 2), and the public sector (i.e., government administration, labelled sector 3).

### 2.1. Manufacturing Sector

We model wage, price and productivity in natural logarithms. We let $w_{1 t}$ denote the wage (per man hour), $q_{1 t}$ denotes the producer price index (in domestic currency) and $z_{1 t}$ denotes labour productivity in fixed prices.

We also measure the national unemployment rate in $\log$ scale, and denote it by $u_{t}$. However, to represent the potential impact of EU labour market integration on manufacturing sector wages we also include the variable $I M$ representing the immigration as a percent of the working population. This variable has not been log-transformed.

In terms of these variables, a long-run wage equation for manufacturing can be written as:

$$
\begin{equation*}
w_{1 t}-q_{1 t}-z_{1 t}=\mu_{w 1}+\beta_{u 1} u_{t}+\beta_{M 1} I M_{t}+e_{1 t}, \beta_{u 1} \leq 0, \beta_{M 1} \leq 0 \tag{1}
\end{equation*}
$$

see e.g., Nickell and Andrews (1983), Hoel and Nymoen (1988), Forslund et al. (2008) among others. $e_{1 t}$ is an unobservable error-term. A typical feature of the data set is that there are dominating positive trends in $q_{1 t}$ and $z_{1 t}$.

For clarity we assume that $q_{1 t}$ and $z_{1 t}$ are random walks with drifts:

$$
\begin{align*}
& q_{1 t}=\mu_{q 1}+q_{1 t-1}+v_{q_{1} t}, \mu_{q 1}>0  \tag{2}\\
& z_{1 t}=\mu_{z 1}+z_{1 t-1}+v_{z_{1} t}, \mu_{z 1}>0 \tag{3}
\end{align*}
$$

where $v_{q_{1} t}$ and $v_{z_{1} t}$ are taken to be two Gaussian processes, for the sake of simplicity. (2) and (3) imply that $q_{1 t}$ and $z_{1 t}$ are integrated of order 1, denoted $I(1)$. Below, we assume that also $u_{t}$ and $I M_{t}$ are $I(1)$.

Given these assumptions, $e_{1 t}$ can logically be either $I(0)$ or $I(1)$. Earlier empirical models of manufacturing wage formation, some of them using data going back to the 1960s, have provided supportive evidence of cointegration, i.e., $I(0)$, see e.g., Nymoen (1989a), Johansen (1995a), Nymoen
and Rødseth (2003). However, none of these studies included the labour immigration variable, $I M$, since the data sets used in these papers ended before EU labour marked enlargement took place.

Equation (1) has interesting implications for wage coordination and pattern wage bargaining. First, since the relationship only involves trending variables from inside the manufacturing sector (i.e., $q_{1 t}$ and $z_{1 t}$ ), it gives a sustainable wage evolution for manufacturing. For wage coordination in particular, the exclusion of wages in the other sectors of the economy is important, as otherwise relative wage effects could damage the profitability and investments in the manufacturing sector, and lack of coordination could lead to wage-wage spirals in the economy.

Second, in the case of cointegration (1) implies that $e_{1 t}$ is an equilibrium correction variable which should logically predict $\Delta w_{1 t+1}$. Third, for manufacturing to act as a wage-leader, the equilibrium correction variable $e_{1 t}$ should also predict $\Delta w_{2 t+1}$ and $\Delta w_{3 t+1}$. The predictive power of $e_{1 t}$ can be tested in for example a cointegrated VAR (CVAR), or in a simultaneous equation model (SEM). In the CVAR, with $\Delta w_{i t}, i=1,2,3$ as left hand side variables, it implies that the loading coefficients of $e_{1 t-1}$ should be non-zero in each of the three rows. In the identified SEM framework we use below, the implication is that $e_{1 t-1}$ should have a non-non zero coefficient in the $\Delta w_{1 t}$ equation, and that $\Delta w_{1 t}$ should have non zero coefficients in the equations of $\Delta w_{2 t}$ and $\Delta w_{3 t}$.

### 2.2. Private Service Sector and Public Sector

Potential long-run relationships for the private service sector wage, $w_{2 t}$, and the government sector wage, $w_{3 t}$ are:

$$
\begin{align*}
& w_{2 t}-w_{1 t}=\mu_{w 2}+\beta_{u 2} u_{t}+\beta_{M 2} I M_{t}+e_{2 t}, e_{2 t} \sim I(0),  \tag{4}\\
& w_{3 t}-w_{2 t}=\mu_{w 3}+\beta_{u 3} u_{t}+\beta_{M 3} I M_{t}+e_{3 t}, e_{3 t} \sim I(0) . \tag{5}
\end{align*}
$$

Conversely, if we find empirically that $e_{2 t}, e_{3 t}$ are $I(1)$, this particular wage pattern does not hold in the data.

In a clear cut pattern wage bargaining model, with manufacturing in the role as the wage leader, the equilibrium correction variable $e_{2 t-1}$ should predict $\Delta w_{2 t}$ and $\Delta w_{3 t}$, but not $\Delta w_{1 t}$. Likewise, $e_{3 t-1}$ should predict $\Delta w_{3 t}$, but not $\Delta w_{1 t}$ and $\Delta w_{2 t}$. In a SEM framework, the implication is that the matrix of contemporaneous coefficients is (lower) diagonal. Subject to (exact) identification assumptions, stated below, these restrictions are testable.

Patterns that are less "clean" than this can also be interpreted as consistent with broad sense wage leader-followership coordination. For example, instead of (5), $w_{3 t}$ may cointegrate directly with $w_{1 t}$. Additionally, in the SEM formulation we can allow a not completely diagonal matrix, for example the coefficient of $\Delta w_{3 t}$ in the $\Delta w_{2 t}$ equation may be non-zero (allowing wage-wage effects between the two sectors), without compromising the wage-leadership of the manufacturing sector.

### 2.3. Wage-Price Inflation

Although the focus of our investigation is the pattern of wage adjustments, it cannot be seen as completely separate from development of consumer prices. In real world wage settlements, demands for wage growth to compensate cost-of-living increases are always on the negotiating table. Since CPI inflation, in turn, depends on growth in wage costs, it is probably unrealistic to have a model of $\Delta w_{1 t}, \Delta w_{2 t}, \Delta w_{3 t}$ and $\Delta p_{t}$ which is recursive. In the short run, nominal wage adjustment also in the wage-leading manufacturing sector is very likely to depend on changes in CPI. For this reason, we attempt to include a model of the logarithm of CPI, and its change (i.e., inflation) in our multiple equation system.

Norway is a small open economy and this also affects CPI inflation. Therefore, an import price index, $p i_{t}$ in log form, is included in the data set. Since import prices are in domestic currency, the inclusion of $p i_{t}$ represents a second channel through which the market for foreign exchange can influence wage-price inflation in our framework, the first being through the wage scope variable.

The empirical relevance of this kind of "imported inflation", together with domestic wage adjustments, has been demonstrated in earlier studies, see e.g., Bårdsen et al. (2005, chp. 8).

More recently, Choo and Kurita (2015) using multivariate cointegration analysis, found that disequilibrium in the money marked have contributed to inflation impetus in Norway. Hence, while it is beyond the scope of this paper, a more consolidated empirical assessment of the different dimensions of inflation dynamics is an interesting topic for further research.

## 3. Data Set and VAR Formulation

We use time series for three wage rates, as well as price and productivity indices. As mentioned, it is conceivable that targets for wage shares and relative wages shift when the labour market fundamentals change. We include the labour immigration rate and the official unemployment rate as indicators of such structural determinants. The definitions of the variables included in the investigation are ${ }^{1}$ :
$W_{i t}$ : index for hourly wage in sector $\mathrm{i}=1,2,3$.
$P_{t}$ : consumer price index.
$Q_{1 t}$ : price deflator of gross value added, manufacturing industry.
$\mathrm{Z}_{1 t}$ : labour productivity, output per hour in manufacturing.
$P I_{t}$ : price deflator of imports of goods and services.
$U_{t}$ : unemployment rate, in per cent, civilian unemployment.
$I M_{t}$ : immigration from EU/EFTA countries, North America, Australia and New Zealand and non-EU Eastern Europe, in per cent of the population aged 15-74.

As mentioned above, lower case letters refer to the logarithm of the original variables. For example, $u_{t}=\log \left(U_{t}\right)$ denotes the log of the unemployment rate. Variables in first differences are denoted by $\Delta$. Subscripts denote time period. For example, $p_{t-4}$ refers to the log of the price level four periods back.

Figure 1 shows the time series for manufacturing wage, value added price index and productivity. It is clearly possible to imagine that these series can contain both unit-root trends and deterministic trends, possibly with breaks. As noted above, in this paper we attempt to model them as $I(1)$ trends with unrestricted drift terms. When testing for cointegration we allow for deterministic trends and for regime shifts in the form of step-dummies. The time plots of the other two key variable, $u_{t}$ and $I M_{t}$ are shown in Figure 2.

We aim to model wage per man hour in three sectors and the domestic price level. The first differences of the four variables are shown in Figure 3. The empirical mean of CPI inflation was higher during the 1990s and later. The same evolution is seen in the wage change data, although less markedly, and more clearly for $\Delta w_{1 t}$ and $\Delta w_{2 t}$ than for $\Delta w_{3 t}$. All four series are characterized by short-run oscillations (negative autocorrelation), mainly due to seasonality. However, the short-run variation is not constant. The CPI-inflation graph indicates heteroscedasticity: compared to the 1980s, the variance of inflation went down became reduced during the 1990s before increasing again early in the new millennium. Heteroscedasticity is also detectable for wage changes. The wage growth variances were lower during the 1990s than either before or since. Indeed, a four-quarter moving average of the four variables would remove seasonality and would probably indicate that they are not $I(1)$, suggesting an $I(2)$ model for further research. ${ }^{2}$

[^60]

Figure 1. Trends in manufacturing wage formation: hourly wage costs $w_{1 t}$, average labour productivity $z_{1}$ and value-added deflator $q_{1}$. All three variables are measured in natural logarithms. The graphs have been scale adjusted for easier comparison.


Figure 2. Plot of the immigration flow (IM) measured as gross immigration in per cent of the working-age population and the natural logarithm of the unemployment rate ( $u$ ). The graph for $u$ has been scale adjusted for easier comparison.


Figure 3. Time plots of the first differences of three wage variable and the consumer price index.
The vector of modelled log-level variables is denoted by $\mathbf{Y}_{w p t}=\left(w_{1 t}, w_{2 t}, w_{3 t}, p_{t}\right)^{\prime}$. We write the VAR as:

$$
\begin{equation*}
\Delta \mathbf{Y}_{w p t}=\sum_{i=1}^{3} \boldsymbol{\Gamma}_{i} \Delta \mathbf{Y}_{w p t-i}+\Pi \mathbf{Y}_{t-1}+\mathbf{Y} \mathbf{Q}_{t}+\varepsilon_{t} \tag{6}
\end{equation*}
$$

where all variables in the cointegration space are included in the extended data vector $\mathbf{Y}_{\mathbf{t}}=$ $\left(\mathbf{Y}_{w p t}, x_{1 t}, u_{t}, I M_{t}, z_{1 t}, p i_{t}, D_{t}\right)^{\prime}$.

In $\mathbf{Y}_{t}$, we have introduced $x_{1 t}$ as notation for the wage scope variable, i.e., $x_{1 t}=q_{1 t}+z_{1 t}$. The appearance of the import price index $\left(p i_{t}\right)$ is a consequence of the decision to model the CPI level as well as wages. As already noted, $u_{t}$ and $I M_{t}$ are also treated as I(1).
$D_{t}$ in $\mathbf{Y}_{t}$ denotes a vector of components that affect the level of $\mathbf{Y}_{w p t}$. In the unrestricted case it includes a constant, a trend and two dummies for regimes: $D_{1 t}$ is an incomes policy dummy which is one from 1988q1 to 2002q4, and zero elsewhere, see Bowitz and Cappelen (2001) and Nymoen (2017). $D_{2 t}$ represents the international financial crisis, and is one in the sub-sample 2009q1-2014q1, and zero elsewhere. We added the two variables into the cointegration relations (i.e., added them to the set of so called restricted variables). We considered using a third dummy, for the monetary policy regime shift in 2001q2, but we found little in support for its inclusion in the model. This decision seems to be validated by the results of the robustness tests mentioned in Section 5.
$\mathbf{Q}_{t}$ represents differences of the variables in $\mathbf{Y}_{t-1}$, as well as centered seasonal dummies and impulse indicators for outliers, notably for 1986q3 (industrial conflict and devaluation) and the first quarters of 2011-2014 (breaks in National accounts).

Table 1 shows a representative test-battery, with p-values in round brackets. $F_{A R}$ is the F-test for autoregressive residuals (order 1-5), see Harvey (1990, pp. 174, 278), Kiviet (1986); F ${ }_{\text {ARCH }}$ tests for ARCH of order 1-4, Engle (1982); $\chi_{N O R M}^{2}$ is the test for departures from normality, Jarque and Bera (1980).

Table 1. Diagnostic tests of VAR residuals, with $p$-values (in parentheses). Sample 1980(1)-2014(4).

|  | $w_{1}$ | $w_{2}$ | $w_{3}$ | $p$ |
| :--- | :---: | :---: | :---: | :---: |
| $F_{A R}$ | $1.49(0.21)$ | $1.61(0.17)$ | $1.53(0.19)$ | $1.46(0.21)$ |
| $F_{A R C H}$ | $0.56(0.21)$ | $0.55(0.70)$ | $0.17(0.95)$ | $0.46(0.76)$ |
| $\chi_{\text {NORM }}^{2}$ | $1.12(0.57)$ | $0.46(0.79)$ | $5.30(0.07)$ | $0.14(0.93)$ |

## 4. Cointegration Analysis

The focus in the cointegration analysis is the $\Pi$ matrix with dimension $4 \times 11$. Let $r$ denote the rank of $\Pi$. Since $x_{1}, u, I M, p i$ and $z_{1}$ are non-modelled variables, $r$ can be $0,1,2,3$ or $4 . r=0$ corresponds to no cointegration, which would imply rejection of the theory we formulated above. If, on the other hand, $r=4$, the variables $x_{1 t}, u_{t}, I M_{t}, z_{1}$, and $p i_{t}$ represent five common trends in the three nominal wage rates and the price index. The theory of wage leadership points to the wage scope variable $x_{1 t}=q_{1 t}+z_{1 t}$ as one dominant trend in the system, and pi ("imported inflation") as another.
$\Pi$ can be written as $\Pi=\alpha \boldsymbol{\beta}^{\prime}$, where $\boldsymbol{\alpha}$ is $4 \times r$ and $\beta(11 \times r)$ is the matrix with cointegration parameters. The wage system outlined above, with a long-run wage setting equation in manufacturing, and two normal wage relativities, is consistent with $r=3$. A decision about $r=4$ may be interpreted as a potential long-run relationship between the consumer prices index and, for example, an import price index and domestic unit labour costs.

Regarding structural breaks, we adopt a broken trend interpretation with three regimes. This seems to be a reasonable model interpretation and representation of the two dummies $D_{1 t}$ and $D_{2 t}$ that were introduced above. This approach also allows ut to make use of the recent advance made by Kurita and Nielsen (2019) in the testing of cointegration rank in partial systems with deterministic terms.

The result of the trace tests for cointegration rank is given in Table 2, together with $5 \%$ critical values that are due to the work of Kurita and Nielsen (2019, Table A1), who have provided code for simulation of the asymptotic distributions of the trace statistics for specifications such as ours. ${ }^{3}$ The values of the trace test statistic are well above the corresponding critical values for test of zero and one cointegration relationships, which supports that the number of stationary long-run relationships may be assumed to be at least two. The third row shows that the hypothesis $r=2$ fails to be rejected.

Table 2. Tests of cointegration rank. Sample 1980(1)-2014(4).

| Eigenvalue <br> $\left(\boldsymbol{\lambda}_{\boldsymbol{i}}\right)$ | $\boldsymbol{H}_{\mathbf{0}}$ | $\boldsymbol{H}_{\mathbf{1}}$ | Trace Test <br> Test Statistics | Critical Values |
| :---: | :---: | :---: | :---: | :---: |
| 0.49 | $r=0$ | $r>0$ | 234.32 | 166.63 |
| 0.41 | $r=1$ | $r>1$ | 139.80 | 121.65 |
| 0.29 | $r=2$ | $r>2$ | 65.34 | 80.315 |
| 0.12 | $r=3$ | $r>3$ | 17.26 | 41.97 |

Endogenous variables: $w_{1}, w_{2}, w_{3}, p$. Restricted variables: $x_{1}=\left(q_{1}+z_{1}\right), u, I M, p i, z$, Trend, $D_{1}, D_{2}$. Unrestricted: Constant $95 \%$ quantiles approximated by the Gamma distribution, simulated by using the Ox code provided with Kurita and Nielsen (2019).

Multiple cointegration relationships are unidentified even after normalization on one endogenous variable in each equation. Identification is therefore relative to theory, which in our case is the system of long-run relationships in Section 2. In Table 3, an (over) identified long-run system is shown. For completeness, we have chosen to continue with $r=4$ to investigate whether also the full set

[^61]potential long-run relationships in our model can be estimated with economically meaningful results, even though the third and fourth relationship do not have the same formal test support as the first two. ${ }^{4}$

The second and third relationships exclude the wage scope variable $x_{1}$. In order to separate the second from the third relationship, we exclude $w_{1}$ and $I M$ from the third, and $w_{3}$ from the second relationship. The three identified wage relationships show a close correspondence with the theory of national wage regulation that we presented above. Finally, the fourth equation in Table 3 represents a long-run price equation. As the last line in Table 3 shows, the over-identifying restrictions cannot be rejected at the $11 \%$ level.

Table 3. Estimated long-run relationships. Sample 1980(1)-2014(4).

$$
\begin{aligned}
& w_{1}=\underset{(.)}{1} x_{1} \underset{(0.11)}{-0.71} u \underset{(0.66)}{-1.91} I M \underset{(0.08)}{+0.07} D_{1}+\underset{(0.15)}{+0.05} D_{2} \\
& w_{2}=\underset{(.)}{1} w_{1} \underset{(0.03)}{-0.17} u \underset{(0.18)}{-0.35} I M \underset{(0.03)}{+0.03} D_{1} \underset{(0.05)}{0.03} D_{2} \\
& w_{3}=\underset{(.)}{1} w_{2} \underset{(0.01)}{-0.03} D_{1} \underset{(0.01)}{0.04} D_{2} \\
& p_{t}=\underset{(0.06)}{0.45} w_{2}-\underset{(0.14)}{0.09} z_{1}+\underset{(0.06)}{0.30} p i \underset{(0.02)}{0.03} D_{1}-\underset{(0.03)}{0.07} D_{2} \\
& \text { Test of over-identifying restrictions: } \chi^{2}(13)=19.45[0.11]
\end{aligned}
$$

The coefficients of the wage relationships are also economically interpretable. In the manufacturing sector, the coefficient $\left(\beta_{u 1}\right)$ with respect to unemployment is somewhat higher in absolute value than the estimates in Nymoen (1989b) and Johansen (1995a), for example, who used data from the 1980s. In private service the numerical significance of the estimated coefficient of $u$ is much smaller. In the public sector the estimated coefficient was so small that it was practical to set it to zero. However, in both sectors, the implied wage responsiveness to unemployment will be inherited from wage setting in manufacturing, through the equations of the system.

The two coefficients of the immigration rate (IM) imply that increased labour supply has affected the nominal wage path negatively. Clearly, since the wave of European labour immigrants only arrived late in our sample, after 2005, these estimates must be interpreted with care, perhaps as a mere "first generation" of estimates of euro-area labour immigration effects on aggregate wages. Nevertheless they represent evidence that the system of wage formation has not been immune to EU labour market enlargement. According to the results, the rather large shift upward in $I M$ has reduced the level of manufacturing sector wage level in particular.

The estimated CPI equation, the fourth line of the table, is interpretable as well. The coefficients of the wage and productivity variables are correctly signed, and the elasticities of wage and import price are realistic, see Bårdsen et al. (2003). The additional restriction of long-run price homogeneity of degree one changed the estimated coefficients of the wage relationships very little. However, that homogeneity restriction was rejected statistically, and therefore we did not impose it on the models we report below.

Conditional on the identified long-run relationships, we can test for the minimum restrictions that imply no feed-back from $w_{2}$ and $w_{3}$ to $w_{1}$. The increase in the test of the over-identifying restrictions is small. Therefore it is not surprising that the three restrictions, $\alpha_{w 1,2}=\alpha_{w 1,3}=0$, are statistically acceptable and the test for the over-identifying restrictions is $\chi^{2}(2)=19.45[0.11]$ (the incremental test yields $\chi 2(2)=2.61[0.27]$ ).

We now turn to the results of the SEM parameterization of the cointegrated VAR to investigate the short run properties of the wage-leader/follower model. The SEM represents a relevant framework

[^62]for analysis of pattern bargaining. In particular, we can test the system of pattern wage bargaining without assuming a particular ordering from the outset.

## 5. Testing the Wage-Leading Hypothesis within a SEM

We can interpret the cointegrated VAR as the reduced form of the SEM:

$$
\begin{equation*}
\mathbf{B}_{0} \Delta \mathbf{Y}_{w p t}=\sum_{i=1}^{3} \mathbf{B}_{i} \Delta \mathbf{Y}_{w p t-i}+\mathbf{A E C} \mathbf{C}_{t-1}+\mathbf{\Psi} \mathbf{Q}_{t}+\boldsymbol{\epsilon}_{t} \tag{7}
\end{equation*}
$$

hence the elements in the diagonal of the contemporaneous coefficient matrix $\mathbf{B}_{0}$ are equal to one. $\mathbf{E C}{ }_{t}$ is the vector which has as elements the equilibrium correction variables defined in accordance with (3):

$$
\begin{align*}
& \hat{e}_{1 t}=w_{1 t}-x_{t}+0.71 u_{t}+1.91 I M_{t}-0.06 D_{1 t}-0.05 D_{2 t}  \tag{8}\\
& \hat{e}_{2 t}=w_{2 t}-w_{1 t}+0.17 u_{t}+0.35 I M_{t}-0.003 D_{1 t}+0.003 D_{2 t}  \tag{9}\\
& \hat{e}_{3 t}=w_{3 t}-w_{2 t}+0.03 D_{1 t}-0.04 D_{2 t}  \tag{10}\\
& \hat{e}_{4 t}=p_{t}-0.45 w_{2 t}+0.09 z_{1 t}-0.30 p i-0.03 D_{1 t}+0.07 D_{2 t} \tag{11}
\end{align*}
$$

The economic theory of wage pattern bargaining has nothing to say about the covariance matrix of SEM disturbances, call it $\Omega$. Hence, in order to be of empirical relevance, that theory must be shown to be identified (or not) for the case of unrestricted $\Omega$. In our study, exact identification of the SEM is therefore based on the specification that $\mathbf{A}$ is diagonal. From that starting point, several specific hypotheses about wage leader/followeship are testable as overidentifying restrictions on $\mathbf{B}_{i}$. In the following we focus mainly on restriction on $\mathbf{B}_{0}$.

In Table 4 we report the FIML estimates of $\mathbf{B}_{0}$, and $\mathbf{A}$, after imposing over-identifying restrictions on the $\mathbf{B}_{i}$ and $\boldsymbol{\Psi}$ matrices. These restrictions aid the estimation of the focus parameters, and are statistically valid, as the test statistic of the over-identifying restrictions, denoted $\chi_{E N C-V A R}^{2}$ at the bottom of the table, shows, cf. Doornik and Hendry (2018).

Table 4. Wage-price simultaneous equations model (SEM). FIML estimates of the contemporaneous coefficient matrix $\hat{\mathbf{B}}_{0}$ (first four columns with estimates) and adjustment coefficients $\hat{\mathbf{A}}$ (last four columns). Standard errors in brackets below the estimates. Sample period 1980(1)-2014(4).

|  | $\Delta w_{1 t}$ | $\Delta w_{2 t}$ | $\Delta w_{3 t}$ | $\Delta p_{t}$ | $\widehat{e}_{1 t-1}$ | $\widehat{e}_{2 t-1}$ | $\widehat{e}_{3 t-1}$ | $\widehat{e}_{4 t-1}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\Delta w_{1 t}$ | -1 | $\begin{gathered} -0.28 \\ (0.20) \end{gathered}$ | $\underset{(0.13)}{-0.11}$ | $\begin{aligned} & 0.44 \\ & (0.21) \end{aligned}$ | $\underset{(0.01)}{-0.03}$ | 0 | 0 | 0 |
| $\Delta w_{2 t}$ | $\begin{aligned} & 0.21 \\ & (0.06) \end{aligned}$ | -1 | $\underset{(0.09)}{-0.19}$ | $\begin{aligned} & 0.42 \\ & (0.13) \end{aligned}$ | 0 | $\underset{(0.06)}{-0.11}$ | 0 | 0 |
| $\Delta w_{3 t}$ | $\begin{gathered} 0.27 \\ (0.11) \end{gathered}$ | $\frac{-0.09}{(0.16)}$ | -1 | $\begin{aligned} & 1.10 \\ & (0.22) \end{aligned}$ | 0 | 0 | $\underset{(0.03)}{-0.20}$ | 0 |
| $\Delta p_{t}$ | ${ }_{(0.05)}^{-0.05}$ | $\begin{aligned} & 0.19 \\ & (0.08) \end{aligned}$ | $\begin{aligned} & 0.07 \\ & (0.06) \end{aligned}$ | -1 | 0 | 0 | 0 | $\underset{(0.01)}{-0.10}$ |
| $\chi_{E N C-V A R}^{2}=74.48[0.06]$ |  |  |  |  |  |  |  |  |

Table 4 shows that all the elements of the $\hat{\mathbf{A}}$ matrix (in the last four columns) with estimated coefficients for the equilibrium correction terms are statistically significant. These estimates are consistent with the conclusions of the cointegration analysis in the previous section.

The estimates of the contemporaneous parameters in $\hat{\mathbf{B}}_{0}$ show a clear pattern: In the row for $\Delta w_{1 t}$, the estimated coefficients of $\Delta w_{2 t}$ and $\Delta w_{3 t}$ are negative ("wrong sign"), but they are statistically insignificant. Conversely, $\Delta w_{1 t}$ has sizeable and significant coefficients in the rows for $\Delta w_{2 t}$ and $\Delta w_{3 t}$, which supports the hypothesis that the manufacturing sector is wage-leading, with private service production and the public sector as wage-followers.

The column with estimated coefficients of $\Delta p_{t}$ gives an indication of the contemporaneous compensation of cost-of-living increases in all three sectors. The estimates are statistically significant. In addition, all though not shown in the table, the wage equations contain significant lags of CPI inflation, in addition to $\Delta p_{t}$ (see Gjelsvik et al. (2015) for details). In sum, the model confirms the role of cost of living consideration in wage formation.

The last row of Table 4, with the results for the inflation equation $\Delta p_{t}$, shows that only $\Delta w_{2 t}$ enters significantly and that increased wages increase prices. This implies that there are within-quarter effects of increased wage cost in private service production to consumer price inflation. In addition, since $\widehat{e}_{4 t}$ includes the domestic wage level with a sizeable weight, the model is consistent with monopolistic price setting. Hence, through private service production and equilibrium correction, the consumer price index is influenced by both domestic wages and the price on imports in domestic currency.

Table 5 contains test statistics for the different pattern bargaining hypotheses that can be formulated as restrictions on the $\hat{\mathbf{B}}_{0}$ matrix. The first two rows in Table 5 test in various aspects whether sector 1 is wage leading the other two sectors. The first row shows three versions of the null hypothesis that manufacturing is not wage leading one or both of the two following sectors. Each test is statistically significant at the $1 \%$ level. In the second line of tests, the second entry supports the assumption that neither wages in the private service nor those in the public sector enter significantly into the identified SEM equation for the manufacturing wage (sector 1). Therefore, the first two rows of table 5 support the wage leadership of sector 1 in the pattern wage bargaining. The third row illustrates the relationship between sectors 2 and 3 . The results show that there is a less clear-cut short-run relationship between wages in the private service sector and the public sector.

Table 5. Likelihood-ratio tests of wage leader/follower restrictions on the model in Table 4.

| Restrictions: | Sec $1 \nrightarrow$ Sec 2 | Sec $1 \nrightarrow \operatorname{Sec} 3$ | Sec $1 \nrightarrow$ Sec 2 and 3 |
| :--- | :---: | :---: | :---: |
|  | $\chi^{2}(1)=11.91^{* *}$ | $\chi^{2}(1)=16.78^{* *}$ | $\chi^{2}(2)=26.23^{* *}$ |
| Restrictions: | Sec $2 \nrightarrow$ Sec 1 | Sec $3 \nrightarrow \operatorname{Sec} 1$ | Sec 2 or $3 \nrightarrow \operatorname{Sec} 1$ |
|  | $\chi^{2}(1)=2.04$ | $\chi^{2}(1)=0.62$ | $\chi^{2}(2)=2.75$ |
| Restrictions: | $\operatorname{Sec} 2 \nrightarrow \operatorname{Sec} 3$ | $\operatorname{Sec} 3 \nrightarrow \operatorname{Sec} 2$ | $\operatorname{Sec} 2 \nrightarrow \operatorname{Sec} 3$ |
|  | $\chi^{2}(1)=0.32$ | $\chi^{2}(1)=4.41^{*}$ | $\chi^{2}(2)=5.90$ |
|  | $*$ and $^{* *}$ denotes significance at the $5 \%$ and $1 \%$ levels. |  |  |

Table 6 shows the results of a restricted estimation where we have imposed wage-leadership on the contemporaneous matrix $\hat{\mathbf{B}}_{0}$ discussed above. Compared to Table 4, the remaining estimated coefficients therefore change only little.

Table 6. Wage-price SEM with wage leader-follower restrictions imposed. FIML estimates of $\hat{\mathbf{B}}_{0}$ and $\hat{\mathbf{A}}$ in Equation (7) with standard errors in brackets below the estimates. Sample period 1980(1)-2014(4).

|  | $\Delta w_{1 t}$ | $\Delta w_{2 t}$ | $\Delta w_{3 t}$ | $\Delta p_{t}$ | $\widehat{e c}_{1 t-1}$ | $\widehat{e c}_{2 t-1}$ | $\widehat{e c}_{3 t-1}$ | $\widehat{e c}_{4 t-1}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\Delta w_{1 t}$ | -1 | 0 | 0 | $\begin{aligned} & 0.30 \\ & (0.16) \end{aligned}$ | $\frac{-0.02}{(0.005)}$ | 0 | 0 | 0 |
| $\Delta w_{2 t}$ | $\begin{aligned} & 0.17 \\ & (0.06) \end{aligned}$ | -1 | 0 | $\begin{aligned} & 0.33 \\ & (0.11) \end{aligned}$ | 0 | $\underset{(0.02)}{-0.14}$ | 0 | 0 |
| $\Delta w_{3 t}$ | $\begin{aligned} & 0.20 \\ & (0.08) \end{aligned}$ | 0 | -1 | $\begin{aligned} & 1.19 \\ & (0.24) \end{aligned}$ | 0 | 0 | $\underset{(0.07)}{-0.26}$ | 0 |
| $\Delta p_{t}$ | 0 | $\begin{aligned} & 0.14 \\ & (0.05) \end{aligned}$ | 0 | -1 | 0 | 0 | 0 | $\underset{(0.01)}{-0.10}$ |
| $\chi_{E N C-V A R}^{2}=86.42[0.03]$ |  |  |  |  |  |  |  |  |

As already noted there are several dummies in the model. However, none of them are closely related to change in monetary policy regime that took place in 2001(2), or to the impact that the EU labour market enlargement might have had on wage regulation.

When a step-dummy for the a monetary policy is added to the price equation, the $t$-value is -0.09 and therefore insignificant. However, inflation targeters anticipated that also wage formation would undergo a change after the new monetary policy regime was introduced, cf. Norges Bank (2002). We find no evidence of this kind of structural break in our investigation: the test of joint significance of the monetary policy change in the three wage equations became $\chi^{2}(3)=1.96$ with $p$-value 0.58 . None of the individual $t$-values were significant. ${ }^{5}$

## 6. Conclusions

We formulated a theoretical model of nominal wage regulation in an open economy. The interpretation of the model as a mechanism of national wage coordination is simple enough: If the wage growth in the manufacturing sector leads the wage setting in the other sectors, the functional income distribution will be relatively stable both in the wage-leading and in the wage-following sectors. The wage relativities between sectors will also be stable.

The empirical testing of the theoretical wage regulation model involves several important steps: Decision of cointegration rank in a partial system, identification of multiple long-run relationships and specification of a dynamic econometric model that allows the testing of alternative hypotheses about the interactions between wage changes in the different sectors of the model. The empirical model must allow for the typical broad sense non-stationarity of the data.

The formal tests of cointegration rank allowed us to reject the hypothesis that there was no cointegration in the system, and gave formal support to two long-run relationships. Tests of over-identifying restriction on the long-run equations supported the interpretation of one of the relationships as a wage curve equation of the manufacturing sector, and of the other as a long-run relativity between the manufacturing wage level, and the wage level in rest of the business sector. In order to complete the model, we chose to include also two other potential long-run relationships. One between wages in the government sector wages and the private sector and another relationship that we identify as a long-run price level equation.

The main conclusion of the testing of alternative hypotheses about pattern wage bargaining within the empirical model, was that the manufacturing sector has operated as a wage leader, and with corresponding clear roles of construction, private service production and retail, and the government sector as wage followers. The relationships are however moderated by labour market indicators. First, we obtain coefficients of the rate of unemployment with expected signs. Second, the econometric results support the hypothesis that the long term wage level may have been reduced as a consequence of immigration made possible by EU labour market enlargement in 2004. However, care must be taken since our results may be affected by an composition effect, i.e., if immigrants have replaced natives in the lowest paid jobs, the average hourly wage may become lowered. Hence, our results are not conclusive evidence of reduced capacity of collective action as a result of increased labour immigration during the first 15 years of the 2000s. Hopefully, ongoing research where natives' wages are modelled, will contribute towards clarification of this issue, cf. Dapi et al. (2019).

We find no indication that the new Norwegian monetary policy established in 2001 has changed the system of wage formation as we have modelled it. In the light of the evidence, the position taken by Norwegian inflation targeters, that an implied "over-determination" of the new regime would force a structural break in wage pattern bargaining, overstated how invasive the monetary policy change would become, cf. Norges Bank (2002). After all, the target of the new monetary policy was the inflation forecast, not inflation itself.

[^63]Author Contributions: The authors made equal contributions. All authors have read and agreed to the published version of the manuscript.
Funding: This research has received funding from Norges Forskningsråd (Grant number: 236935).
Conflicts of Interest: The authors declare no conflict of interest.

## References

Aukrust, Odd. 1977. Inflation in the Open Economy. A Norwegian Model. In World Wide Inflation. Theory and Recent Experience. Edited by Lawrence Krause and Walter Sâlant. Washington: Brookings.
Bårdsen, Gunnar, Øvind Eitrheim, Eilev S. Jansen, and Ragnar Nymoen. 2005. The Econometrics of Macroeconomic Modelling. Oxford: Oxford University Press.
Bårdsen, Gunnar, Eilev S. Jansen, and Ragnar Nymoen. 2003. Econometric Inflation Targeting. Econometrics Journal 6: 429-60. [CrossRef]
Bårdsen, Gunnar, and Ragnar Nymoen. 2003. Testing steady-state implications for the NAIRU. The Review of Economics and Statistics 85: 1070-75. [CrossRef]
Bowitz, Einar, and Ådne Cappelen. 2001. Modeling income policies: Some Norwegian experiences 1973-1993. Economic Modelling 18: 349-79. [CrossRef]
Calmfors, Lars, and Anna Larsson Seim. 2013. Pattern Bargaining and Wage Leadership in a Small Open Economy. The Scandinavian Journal of Economics 115: 109-40. [CrossRef]
Camarero, Mariam, Gaetano D'Adamo, and Cecilio Tamarit. 2014. Wage leadership models: A country-by-country analysis of the EMU. Economic Modelling 44: S2-S11. [CrossRef]
Camarero, Mariam, Gaetano D'Adamo, and Cecilio Tamarit. 2016. The role of institutions in explaining wage determination in the Eurozone: A panel cointegration approach. International Labour Review 155: 25-56. [CrossRef]
Choo, Han Gwang, and Takamitsu Kurita. 2015. A monetary perspective on inflation dynamics in Norway. International Journal of Economic Research 12: 583-606.
Courbis, Raymond. 1974. Liason IInternational des Pric et Inflation Importé. Economie Appliqué 27: 2015-220.
Cukierman, Alex, and Francesco Lippi. 1999. Central bank independence, centralization of wage bargaining, inflation and unemployment. European Economic Review 43: 1395-434. [CrossRef]
Dapi, Bjorn, Ragnar Nymoen, and Victoria Sparrman. 2019. Robustness of the Norwegian Wage Formation System and Free EU Labour Movement: Evidence from Wage Data for Natives. Discussuion Papers No. 895. Oslo: Statistics Norway.
Doornik, Jurgen A. 2018. An Object-Oriented Matrix Programming Language, Ox 8. London: Timberlake Consultants.
Doornik, Jurgen A., and David F. Hendry. 2018. Modelling Dynamic Systems, PcGive 15. London: Timberlake Consultants, vol. 2.
Driehuis, Wim, and Pieter de Wolff. 1976. A Sectoral Wage Price Model for The Netherlands' Economy. In Inflation in Small Countries. Edited by Helmut Frisch. New York: Springer, pp. 283-339. [CrossRef]
Edgren, Gösta, Karl-Olof Faxén, and Clas-Erik Odher. 1969. Wages, Growth and Distribution of Income. Swedish Journal of Economics 71: 133-60. [CrossRef]
Engle, Robert F. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50: 987-1007. [CrossRef]
Engle, Robert F., and Clive W. J. Granger. 1987. Co-Integration and Error Correction: Representation, Estimation, and Testing. Econometrica 55: 251. [CrossRef]
Evju, Stein. 2014. Norway. In Viking, Laval and Beyond. Oxford: Hart Publishing, pp. 211-28.
Forder, James. 2014. Macroeconomics and the Phillips Curve Myth. Oxford Studies in the History of Economics. Oxford: Oxford University Press.
Forslund, Anders, Nils Gottfries, and Andreas Westermark. 2008. Prices, Productivity and Wage Bargaining in Open Economies. Scandinavian Journal of Economics 110: 169-95.
Gjelsvik, Marit Linnea, Ragnar Nymoen, and Victoria Sparrman. 2015. Have Inflation Targeting and EU Labour Immigration Changed the System of Wage Formation in Norway? Memorandum No 18/2015. Oslo: University of Oslo.
Harvey, Andrew. 1990. The Econometric Analysis of Time Series, 2nd ed. London: Philip Allan.

Hoel, Michael, and Ragnar Nymoen. 1988. Wage Formation in Norwegian Manufacturing. An Empirical Application of a Theoretical Bargaining Model. European Economic Review 32: 977-97. [CrossRef]
Holden, Steinar. 2005. Monetary regimes and the co-ordination of wage setting. European Economic Review 49: 833-43. [CrossRef]
Iversen, Torben. 1999. Contested Economic Institutions. The Politics of Macroeconomics and Wage Bargaining in Advanced Democracies. Cambridge: Cambridge University Press.
Jarque, Carlos M., and Anil K. Bera. 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Economics Letters 6: 255-59. [CrossRef]
Johansen, Kåre. 1995a. Norwegian Wage Curves. Oxford Bulletin of Economics and Statistics 57: 229-47. [CrossRef]
Johansen, Søren. 1988. Statistical Analysis of Cointegration Vectors. Journal of Economic Dynamics and Control 12: 231-54. [CrossRef]
Johansen, Søren. 1995b. Likelihood-Based Inference in Cointegrated Vector Autoregressive Models. Oxford: Oxford University Press. [CrossRef]
Johansen, Søren, and Katarina Juselius. 1990. Maximum likelihood estimation and inference on cointegration-With applications to the demand for money. Oxford Bulletin of Economics and Statistics 52: 169-210. [CrossRef]
Juselius, Katarina. 2007. The Cointegrated VAR Model: Methodology and Applications. Oxford: Oxford University Press.
Kiviet, Jan F. 1986. On the Rigour of Some Misspecification Tests for Modelling Dynamic Relationships. The Review of Economic Studies 53: 241. [CrossRef]
Knell, Markus, and Alfred Stiglbauer. 2009. Wage Staggering and Wage Leadership in Austria. Review and Implications. Monetary Policy and the Economy 4: 79-97.
Kolsrud, Dag, and Ragnar Nymoen. 2014. Macroeconomic Stability or Cycles? The Role of the Wage-Price Spiral. Australian Economic Papers 53: 41-68. [CrossRef]
Kolsrud, Dag, and Ragnar Nymoen. 2015. Heuristic vs formal dynamics of the wage- and price-curve model of equilibrium unemployment. Journal of Economic Studies 42: 186-206. [CrossRef]
Kurita, Takamitsu and Bent Nielsen. 2019. Partial cointegrated vector autoregressive models with structural breaks in deterministic terms. Econometrics 7: 42. [CrossRef]
Lamo, Ana, Javier J. Pérez, and Ludger Schuknecht. 2012. Public or Private Sector Wage Leadership? An International Perspective*. The Scandinavian Journal of Economics 114: 228-44. [CrossRef]
Lindquist, Matthew, and Roger Vilhelmsson. 2006. Is the Swedish central government a wage leader? Applied Economics 38: 1617-25. [CrossRef]
Meade, James E. 1982. Wage-Fixing, Stagflation Crows Nest: George Allen and Unwin, vol. 1.
Nickell, Stephen J., and Michael Andrews. 1983. Unions, Real-Wages and Employment in Britain 1951-79. Oxford Economic Papers (Supplement) 35: 183-206. [CrossRef]
Norges Bank. 2002. Inflation Report 3/2002. Oslo: Norges Bank.
Nymoen, Ragnar. 1989a. Modelling Wages in the Small Open Economy: An Error-Correction Model of Norwegian Manufacturing Wages. Oxford Bulletin of Economics and Statistics 51: 239-58. [CrossRef]
Nymoen, Ragnar. 1989b. Wages and the Length of the Working Day. A Empirical Test Based on Norwegian Quarterly Manufacturing Data. Scandinavian Journal of Economics 91: 599-612. [CrossRef]
Nymoen, Ragnar. 1991. A Small Linear Model of Wage- and Price-Inflation in the Norwegian Economy. Journal of Applied Econometrics 6: 255-69. [CrossRef]
Nymoen, Ragnar. 2017. Between Institutions and Global Forces: Norwegian Wage Formation Since Industrialisation. Econometrics 5: 6. [CrossRef]
Nymoen, Ragnar, and Asbjørn Rødseth. 2003. Explaining Unemployment: Some Lessons from Nordic Wage Formation. newblock. Labour Economics 10: 1-29. [CrossRef]
Sargan, John Denis. 1964. Wages and Prices in the United Kingdom: A Study of Econometric Methodology. In Econometric Analysis for National Economic Planning. Edited by Peter E. Hart, Gordon Mills and John K. Whitaker. London: Butterworth Co., pp. 25-63.

Sargan, John Denis. 1980. A Model of Wage-Price Inflation. The Review of Economic Studies 47: 97. [CrossRef]
Soskice, David. 1990. Wage determination: The changing role of institutions in advanced industrialized countries. Oxford Review of Economic Policy 6: 36-61. [CrossRef]
Soskice, David, and Torben Iversen. 2000. The Nonneutrality of Monetary Policy with Large Price or Wage Setters. Quarterly Journal of Economics 115: 265-84. [CrossRef]
© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http:/ / creativecommons.org/licenses/by/4.0/).

## Article

# Forward Rate Bias in Developed and Developing Countries: More Risky Not Less Rational 

Michael D. Goldberg ${ }^{1, *, \dagger}$, Olesia Kozlova ${ }^{2, \dagger}$ and Deniz Ozabaci ${ }^{1, \dagger}$<br>1 School of Business and Economics, University of New Hampshire, 10 Garrison Avenue, Durham, NH 03824, USA; deniz.ozabaci@unh.edu<br>2 Capital Group, 333 South Hope Street, Los Angeles, CA 90071, USA; olesia.a.kozlova@gmail.com<br>* Correspondence: michael.goldberg@unh.edu; Tel.: +1-603-862-3385<br>$\dagger$ The authors are heavily indebted to the guest editors, Rocco Mosconi and Paolo Paruolo, for their generous help and many insights on improving the paper's empirical analysis. We also thank the referees for useful comments and suggestions. We have also benefited from comments and reactions from Bruce Elmslie, Jay Horvath, Katarina Juselius, and participants of the University of New Hampshire's research seminar. We are grateful to the Institute for New Economic Thinking (INET) and the University of New Hampshire for financial support.

Received: 10 April 2018; Accepted: 3 November 2020; Published: 2 December 2020


#### Abstract

This paper examines the stability of the Bilson-Fama regression for a panel of 55 developed and developing countries. We find multiple break points for nearly every country in our panel. Subperiod estimates of the slope coefficient show a negative bias during some time periods and a positive bias during other time periods in nearly every country. The subperiod biases display two key patterns that shed light on the literature's linear regression findings. The results point toward the importance of risk in currency markets. We find that risk is greater for developed country markets. The evidence undercuts the widespread view that currency returns are predictable or that developed country markets are less rational.


Keywords: imperfect knowledge; Knightian Uncertainty; structural change; currency risky

## 1. Introduction

The forward rate anomaly is a long-standing puzzle in International Macroeconomics. The anomaly is based on a linear regression of the future change of the spot exchange rate on the forward premium. The assumptions of risk neutral investors, no capital controls, and the rational expectations hypothesis (REH) imply that the slope coefficient (hereafter $\beta$ ) should be unity. However, Bilson (1981) and Fama (1984) (BF) and many subsequent studies report that $\beta$ is less than unity and negative in the major currency markets. ${ }^{1}$ The negative bias from unity "suggests that one can make predictable profits by betting against the forward rate" Obstfeld et al. (1996), p. 589.

Macroeconomists have explored two main explanations for the predictable excess returns: a time-varying risk premium or systematic forecasting errors. REH risk premium models have encountered considerable difficulty in explaining the negative bias, which has given support to models in which market participants are less than fully rational. ${ }^{2}$

The cointegrated VAR (CVAR) studies of (Juselius 2017a, 2017b; Juselius and Assenmacher 2017; Juselius and Stillwagon 2018) provide evidence that the forward rate anomaly may originate from another source. These studies find that the process underpinning currency returns is not only unstable,

[^64]but the instability is triggered by novel historical developments such as German reunification and the 1985 Plaza Accord. The instability implies that market participants must cope with imperfect knowledge about the future and that returns are less predictable than widely reported. The CVAR findings also provide evidence of a time-varying risk premium, but one based on imperfect knowledge economics (IKE). ${ }^{3}$ In the IKE model, the market's risk premium compensates participants for their loss aversion and downside risk. The model relates downside risk to the gap between the exchange rate and its benchmark value, rather than to the volatility of returns as with standard REH models. Juselius, Assenmacher, and Stillwagon use purchasing power parity (PPP) to define benchmark values. They find that excess returns are positively related to the gap from PPP at high significance levels as predicted by the IKE model. ${ }^{4}$ Taken as a whole, the CVAR findings suggest that forward rate biasedness may be better understood as a consequence of imperfect knowledge and risk, rather than a lack of rationality.

In this paper, we present additional evidence of this view. Our analysis examines a key finding in the literature: forward rate biasedness is less negative for developing countries than for developed countries. Most studies report that $\beta$, although less than unity, is positive for developing countries. ${ }^{5}$ However, these countries are generally thought to be riskier for investors than developed countries. They are characterized by greater political and macroeconomic instability, less liquid and more volatile financial markets, and greater vulnerability to commodity price and other terms of trade shocks. These countries' currency markets should thus be characterized by larger and more volatile risk premiums and thus greater forward rate biasedness. The finding that they are not is taken by Frankel and Poonawala (2010) and others to imply a striking conclusion: currency markets in developed countries are less rational than those in developing countries. ${ }^{6}$

We argue, however, that this conclusion, and the broader claim of predictable excess returns, misses what is arguably the key problem facing currency forecasters: instability in the process underpinning outcomes (Clements and Hendry (1999)). ${ }^{7}$ We go further than (Juselius 2017a, 2017b); Juselius and Assenmacher (2017), and Juselius and Stillwagon (2018), and others in documenting this structural change. Our structural change analysis is comprehensive; we examine the BF regression's instability for a large panel of 20 developed and 35 developing countries. Our sample of monthly observations runs from the mid 1980s through January 2016 for most developed countries and the late 1990s (due to data availability) through January 2016 for many developing countries.

We find that the BF regression is characterized by multiple structural breaks for nearly all countries in the full sample. The breakpoints for each country, in turn, imply multiple subperiods or "regimes" that are characterized by a distinct $\beta$. In roughly half of all the subperiods for both groups of countries, $\beta=1$ cannot be rejected. In the other subperiods, we find regimes in which $\beta>1$ and other regimes in which $\beta<1$ (and sometimes negative) for nearly all countries. The results show that there are prolonged time periods in which one would have earned profits on average by betting against the

[^65]forward rate $(\beta<1)$ and other time periods in which one would have either earned profits by betting with the forward rate $(\beta>1)$ or earned no profits at all $(\beta=1) .{ }^{8}$

These results undercut the widespread view that currency returns are predictable on the basis of the linear BF regression. To predict returns, one would need to predict the structural change that underpins outcomes in these markets. As such, the literature's linear-regression estimates provide little evidence, one way or the other that currency markets in developed countries are less rational than those in developing countries.

Nonetheless, the sharp difference in the linear-regression estimates for the two groups of countries is intriguing and raises two sets of questions. First, do the subperiod estimates of $\beta$ display patterns that shed light on why the linear estimates show a greater negative bias for developed countries? In addition, second, are the patterns informative of the importance of imperfect knowledge and risk in currency markets? We find affirmative answers to both questions. A key result is that the size of the subperiod biases, negative and positive, are roughly two times larger for developed countries than for developing countries. As such, the structural changes that occur in developed-country markets are considerably larger and thus lead to greater capital losses when structural change occurs.

The remainder of the paper is structured as follows: Section 2 extends Frankel and Poonawala (2010)'s linear-regression analysis by updating the sample period and enlarging the panel to 55 countries. We find that Frankel and Poonawala's main result of a smaller bias for developing countries is weakened in the extended panel. In Section 3, we test for instability in the BF regression for the countries in our panel. We rely on recursive procedures that leave open the timing, magnitude, and number of structural breaks in the data. The section reports two key patterns in the subperiod biases. Sections 4 and 5 discuss how these patterns point toward the importance of imperfect knowledge and risk in currency markets. Section 6 offers concluding remarks.

## 2. Updating the Linear Estimates: Evidence of Instability

The forward rate anomaly is based on the BF regression:

$$
\begin{equation*}
\Delta s_{t+1}=\alpha+\beta f p_{t}+\varepsilon_{t+1} \tag{1}
\end{equation*}
$$

where $s_{t+1}$ denotes the log of the spot exchange rate at time $t+1$ (the domestic currency price of foreign exchange), $f p_{t}$ is the forward premium on foreign exchange (the log of the forward rate minus the log of the spot rate), $\varepsilon_{t+1}$ is an error term, and $\Delta$ is a first-difference operator.

Frankel and Poonawala (2010) consider the BF regression for a panel of 20 developed and 14 developing countries and a sample period that begins in December 1996 for most countries and runs through April 2004. We first reproduce their main results. We then consider a larger panel of 55 countries, consisting of the original 34 countries plus 21 additional developing economies. We also update the sample period so that it begins before December 1996 for most of the developed countries and some of the developing countries and runs through January 2016. ${ }^{9}$ We use monthly data on spot and one-month forward rates for nearly all countries, which we obtain from Thompson DataStream's World Market Reuters (WMR). The one exception is New Zealand, for which WMR does not provide a consistent forward rate series. For this country, we input one-month eurocurrency interest rates and the spot exchange rate into covered interest parity (CIP) to derive a one-month forward exchange rate. ${ }^{10}$

There is clearly dependence in the data. Many of the countries in the original and extended panels had some type of pegged or managed exchange rate regime over most or all of the sample period.

[^66]The panels include countries that were in the European Exchange Rate Mechanism (ERM) until 1998, members of the euro area starting in 1999 and later, and non-European countries, such as Hong Kong and Saudi Arabia, which maintained tight U.S. dollar pegs. However, even for floating-rate regimes, we would expect U.S. macro news to impact all of the U.S. dollar exchange rates in our panel, often in the same direction.

Frankel and Poonawala include pegged-rate and managed-floating regimes in their analysis. They first present OLS results for all developed and developing countries, including the 11 individual euro countries. To account for the dependence in the data, they employ balanced seemingly unrelated regressions (SUR). The data for the euro countries is largely overlapping (from 1999-2004). Frankel and Poonawala thus drop individual euro countries in the SUR analysis and use single euro-area spot and forward rate series for the region. However, they continue to include other tightly pegged developed- and developing-country regimes in the SUR estimation (for example, Denmark, Hong Kong, and Saudi Arabia).

We reproduce Frankel and Poonawala's OLS and main SUR results for their original panel of countries and sample periods. In order to fully exploit our extended data set, we also estimate an unbalanced SUR model that includes all developed and developing countries. ${ }^{11}$ We use single euro-area spot and forward rate series for the 11 developed euro countries in our panel, but, unlike Frankel and Poonawala, we treat Denmark as in the euro-area. ${ }^{12,13}$ Our extended panel includes five developing euro countries that joined after the single currency's inception (Estonia, Latvia, Lithuania, Slovenia, and Slovakia). The samples for the first four of these countries begin in April 2004. By that date, their currencies were (or soon to be) tied to the euro in an ERM. We thus treat these countries like the other euro-area countries and drop their individual exchange rate series from the SUR analysis. Slovakia's sample, however, begins in March 2002. The country had a floating currency until the end of November 2005, after which it tied its koruna to the euro as part of an ERM. We thus include Slovakia as an additional country in our SUR analysis, ending its sample in November 2005.

Frankel and Poonawala's rationale for including some tightly pegged regimes, but not others in their analysis is unclear. Including them either makes sense economically or not.

Pegged rate regimes are characterized by exchange rate fluctuations, albeit in a much more narrow range than managed floating rate regimes. Although small, the fluctuations should be consistent with forward rate unbiasedness under the assumptions of risk neutrality and REH, thereby providing economic rationale for including them. Any incipient deviation from unbiasedness would lead to large expected profits and massive capital flows that would quickly push rates back into line. Moreover, pegged and managed-rate regimes are characterized, on the whole, by smaller PPP deviations than their floating-rate counterparts. According to the IKE model, these regimes should be less prone to downside risk. The question is whether the many developing countries that have them are characterized by lower forward rate biasedness. Excluding these countries, therefore, may bias the analysis against such a finding, hindering attempts to uncover whether risk considerations or irrationality underpinned the pattern of forward rate biasedness across developed and developing countries.

However, the range of fluctuations in pegged rate regimes may be so narrow, and the expected exchange rate changes so small that the expected profits from exploiting deviations may be smaller

[^67]than the transactions costs (given by bid-asked spreads) and cost of capital. In this case, deviations from unbiasedness would not result in capital flows and thus not be a reflection of the importance of risk or irrationality. ${ }^{14}$ The Saudi Arabian riyal (SAR) is a case in point. Except for a few brief periods, the riyal was tightly pegged to the U.S. dollar (USD) at 3.7502 for nearly the entire sample (see Figure 1). The coefficient of variation of the spot rate over our sample is 0.0003 , which is lower than the currency's average bid-asked spread (0.0004) over the sample. ${ }^{15}$ Deviations from unbiasedness in the SAR market (and other tightly pegged regimes), therefore, may not be economically meaningful, thereby weakening the strength of the empirical results.


Figure 1. SAR-USD Spot Exchange Rate 1990:06-2016:01.
In order to be conservative, we conduct the SUR analysis for a panel that also excludes the other tightly pegged regimes, which involve USD pegs. The group entails the following countries/time periods: Saudi Arabia, Hong Kong, Bahrain, China and Malaysia prior to August 2005, and Thailand before August 1997. ${ }^{16}$

[^68]
### 2.1. Frankel and Poonawala's Panel

Tables 1 and 2 (columns 3 and 4) report OLS results for Frankel and Poonawala's original panel of countries and sample periods. ${ }^{17}$ The results suggest that forward rate biasedness is much greater for developed countries. The $\beta$ estimates (denoted by $\hat{\beta}$ ) for developed countries are all negative (save for Greece). In total, 17 of the 20 economies are characterized by negative biasedness at the $5 \%$ level (that is, $\beta<1$ ), which we indicate by bold figures in the table. By contrast, only four of the 14 original developing countries are characterized by a negative $\hat{\beta}$, and only seven of the countries in this group have a negative bias at the $5 \%$ level. The average $\hat{\beta}$ for the group of developed and developing countries is -4.12 and 0.27 , respectively.

Table 1. Linear individual country BF regressions for developed countries.

| Developed Country | Full-Sample Time Period | $\hat{\boldsymbol{\beta}}$ FP (2010) | Rob. SE | $\hat{\boldsymbol{\beta}}$ Full Sample | Rob. SE |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Australia | $12 / 84-01 / 16$ | $-\mathbf{6 . 5 8 9 5}$ | 2.0660 | $-\mathbf{1 . 0 7 9 0}$ | 0.5862 |
| Austria | $01 / 76-01 / 16$ | -5.3837 | 2.1372 | 0.3473 | 0.5671 |
| Belgium | $03 / 85-01 / 16$ | -3.0095 | 2.0691 | $-\mathbf{0 . 1 4 3 0}$ | 0.0881 |
| Canada | $12 / 84-01 / 16$ | $-\mathbf{3 . 1 3 8 0}$ | 1.6270 | $-\mathbf{0 . 5 3 8 5}$ | 0.7371 |
| Denmark | $12 / 84-01 / 16$ | -5.5065 | 2.0821 | -0.0773 | 0.6478 |
| Finland | $01 / 97-01 / 16$ | -5.0479 | 1.5597 | -2.5680 | 1.5296 |
| France | $01 / 97-01 / 16$ | -4.9574 | 2.1393 | $-\mathbf{2 . 3 7 2 6}$ | 1.6716 |
| Germany | $01 / 97-01 / 16$ | $-\mathbf{4 . 9 4 7 7}$ | 2.0923 | $-\mathbf{2 . 3 9 0 6}$ | 1.6420 |
| Greece | $01 / 97-01 / 16$ | 2.8595 | 1.4633 | 1.6651 | 0.8928 |
| Ireland | $08 / 86-01 / 16$ | -5.5840 | 2.2778 | 0.2311 | 0.9550 |
| Italy | $01 / 97-01 / 16$ | $-\mathbf{4 . 1 5 3 6}$ | 2.1424 | -1.8370 | 1.7542 |
| Japan | $10 / 83-01 / 16$ | -1.5469 | 2.0916 | $-\mathbf{1 . 1 3 6 2}$ | 0.8807 |
| Netherlands | $10 / 83-01 / 16$ | $-\mathbf{2 . 9 5 1 4}$ | 2.0315 | $-\mathbf{0 . 5 6 1 8}$ | 0.7746 |
| New Zealand | $01 / 75-01 / 16$ | -7.7074 | 1.9594 | -0.3427 | 1.0636 |
| Norway | $12 / 84-01 / 16$ | -3.4212 | 1.2589 | $-\mathbf{0 . 3 7 9 0}$ | 0.7459 |
| Portugal | $01 / 76-01 / 16$ | -4.6132 | 2.3530 | $\mathbf{0 . 6 5 5 0}$ | 0.1677 |
| Spain | $08 / 86-01 / 16$ | -5.3954 | 2.2810 | $-\mathbf{0 . 9 2 5 1}$ | 0.6121 |
| Sweden | $12 / 84-01 / 16$ | -5.0888 | 1.2093 | $-\mathbf{0 . 1 8 1 8}$ | 1.1258 |
| Switzerland | $10 / 83-01 / 16$ | -3.8778 | 2.2572 | $-\mathbf{1 . 4 2 8 9}$ | 1.0171 |
| UK | $10 / 83-01 / 16$ | -2.3769 | 2.8693 | -0.8198 | 1.5206 |
| Average |  | -4.1219 |  | -0.6941 |  |

Equation (1) is estimated by ordinary least squares (OLS) with Newey-West robust standard errors to correct for possible heteroskedastic and serially correlated errors. Constant terms in the regression are not reported. Frankel and Poonawala (2010) denotes Frankel and Poonawalaâs sample periods, which run from 1996/12 through 2004/04 for most countries. Column 2 shows some countriesâ samples start at â1/97â Consequently, the number of observations for these countries is 87 , whereas for the rest, it is 88 as in Frankel and Poonawala (2010). Figures in bold indicate rejections of the null that $\beta=1$ at the $5 \%$ level, against the one-sided alternative $\beta<1$.

[^69]Table 2. Linear individual country BF regressions for developed countries.

| Developing Country | Full-Sample <br> Time Period | $\hat{\beta}$ Frankel and Poonawala (2010) | Rob. SE | $\begin{gathered} \hat{\beta} \text { Full } \\ \text { Sample-FP } \\ \text { Countries } \end{gathered}$ | $\hat{\beta}$ Full Sample-Non-FP Countries | Rob. SE |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Argentina | 04/04-01/16 |  |  |  | 0.8784 | 0.0766 |
| Bahrain | 05/00-01/16 |  |  |  | -0.0418 | 0.0726 |
| Brazil | 07/00-01/16 |  |  |  | 0.1155 | 0.1471 |
| Bulgaria | 04/04-01/16 |  |  |  | 0.4697 | 1.9372 |
| Chile | 04/04-01/16 |  |  |  | 1.9463 | 1.1836 |
| China | 03/02-01/16 |  |  |  | 0.5732 | 0.1673 |
| Colombia | 03/99-01/16 |  |  |  | 1.4912 | 0.4180 |
| Czech Rep. | 01/97-01/16 | 1.3479 | 1.2396 | 1.0205 |  | 0.9551 |
| Estonia | 04/04-01/16 |  |  |  | -0.9089 | 1.3922 |
| Hong Kong | 10/83-01/16 | 0.0593 | 0.0840 | 0.1259 |  | 0.0522 |
| Hungary | 11/97-01/16 | 1.1647 | 1.3719 | -0.887 |  | 0.7263 |
| India | 11/97-01/16 | -0.8749 | 0.4741 | -0.2024 |  | 0.5131 |
| Indonesia | 01/97-01/16 | 0.2430 | 0.2579 | 0.2541 |  | 0.2479 |
| Israel | 07/98-01/16 |  |  |  | -0.2879 | 0.3174 |
| Kuwait | 06/90-01/16 | 0.6315 | 0.8120 | 1.3529 |  | 0.8039 |
| Latvia | 04/04-01/16 |  |  |  | -1.6891 | 0.8128 |
| Lithuania | 04/04-01/16 |  |  |  | 0.3618 | 1.7082 |
| Malaysia | 11/97-01/16 |  |  |  | -0.0354 | 0.0123 |
| Mexico | 01/97-01/16 | -0.4879 | 0.3188 | -0.1656 |  | 0.2993 |
| Morocco | 04/04-01/16 |  |  |  | 0.3165 | 0.5727 |
| Pakistan | 05/98-01/16 |  |  |  | 0.0792 | 0.0877 |
| Peru | 04/04-01/16 |  |  |  | 0.9109 | 0.4792 |
| Philippines | 01/97-01/16 | 1.1688 | 1.3846 | 1.4306 |  | 0.9806 |
| Poland | 09/96-01/16 |  |  |  | 0.6004 | 0.4437 |
| Romania | 04/04-01/16 |  |  |  | -0.6929 | 1.0140 |
| Russia | 04/04-01/16 |  |  |  | 2.5016 | 0.2242 |
| S. Africa | 06/90-01/16 | -3.3386 | 1.7908 | -1.6162 |  | 1.0938 |
| S. Arabia | 06/90-01/16 | -0.0435 | 0.0265 | -0.0871 |  | 0.0539 |
| S. Korea | 04/98-01/16 |  |  |  | 0.5351 | 0.5648 |
| Slovakia | 03/02-01/16 |  |  |  | -2.7064 | 0.8820 |
| Slovenia | 04/04-01/16 |  |  |  | 1.1925 | 2.2092 |
| Singapore | 12/84-01/16 | 1.1711 | 0.7445 | 0.9057 |  | 0.4659 |
| Taiwan | 01/97-01/16 | 0.8770 | 0.7308 | 0.7942 |  | 0.4810 |
| Thailand | 03/95-01/16 | 1.8896 | 0.3506 | 1.5830 |  | 0.5017 |
| Turkey | 01/97-01/16 | 0.0200 | 0.0348 | -0.0047 |  | 0.0225 |
| Average |  | 0.2734 |  | 0.3172 | 0.2671 |  |

Equation (1) is estimated by ordinary least squares (OLS) with Newey-West robust standard errors to correct
for possible heteroskedastic and serially correlated errors. Constant terms in the regression are not reported.
Frankel and Poonawala (2010) denotes Frankel and Poonawalaâs sample periods, which run from 1996/12
through 2004/04 for most countries. Column 2 shows some countriesâ samples start at â1/97â Consequently,
the number of observations for these countries is 87 , whereas for the rest, it is 88 as in Frankel and Poonawala
(2010). Figures in bold indicate rejections of the null that $\beta=1$ at the $5 \%$ level, against the one-sided alternative
$\beta<1$.
The balanced SUR results in columns 2 and 3 of Table 3 also show greater negative biasedness for developed countries, although less so compared with the OLS estimates. ${ }^{18}$ Nine of the 10 developed country $\hat{\beta}$ s are negative, whereas negative estimates are found for eight of the 13 developing countries. ${ }^{19}$ However, the percentage of developed countries that are characterized by a significant negative bias (again shown in bold in the tables) drops to 60 percent after accounting for data dependence, and we observe a similar percentage for developing countries. A pooled unbalanced SUR model shows that both groups of countries are characterized by negative forward rate bias (see Table 4, column 4). ${ }^{20}$ However, the results in Table 4 (column 3) also show, like Frankel and Poonawala,

[^70]greater biasedness for developed countries: $\hat{\beta}$ for the group of developed countries is significantly less than zero, whereas the zero null cannot be rejected for the developing country group.

Table 3. Seemingly unrelated regression estimates.

| Developed |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
|  | FP-2010 | SE | Full Sample | SE |
| Australia | -0.6561 | 1.6124 | -0.4604 | 0.5696 |
| Canada | -0.6173 | 1.9678 | -0.3907 | 0.5672 |
| Denmark | -2.0294 | 1.0593 |  |  |
| Euro Area | -1.938 | 0.8706 | 0.2353 | 0.4729 |
| Japan | 1.7178 | 1.7425 | -0.5107 | 0.7336 |
| New Zealand | -2.9392 | 1.9216 | -0.3512 | 0.8266 |
| Norway | -1.4489 | 0.8016 | -0.2112 | 0.4926 |
| Sweden | -2.2555 | 1.0263 | 0.0901 | 0.7325 |
| Switzerland | -2.4694 | 1.1320 | -1.2431 | 0.6492 |
| UK | -0.494 | 1.4048 | -0.7479 | 0.9238 |
| Developing |  |  |  |  |
|  |  |  | 0.8454 | 0.1237 |
| Bahrain |  |  |  |  |
| Brazil |  |  | 0.1048 | 0.0552 |
| Bulgaria |  |  | 0.3218 | 1.5762 |
| Chile |  |  | 1.0252 | 0.8838 |
| China |  |  | 0.6377 | 0.1404 |
| Colombia |  |  | 1.2850 | 0.3175 |
| Czech Rep. | -0.3647 | 0.6256 | 0.5578 | 0.7574 |
| Hong Kong | 0.0429 | 0.0775 |  |  |
| Hungary | -0.2275 | 0.6461 | 0.0323 | 0.2987 |
| India | -0.4344 | 0.3633 | 0.2357 | 0.3175 |
| Indonesia |  |  | 0.2687 | 0.1831 |
| Israel |  |  | -0.3092 | 0.3427 |
| Kuwait | 0.7167 | 0.4929 | 1.1272 | 0.4778 |
| Malaysia |  |  | -0.0312 | 0.0204 |
| Mexico | -0.6581 | 0.4038 | 0.0914 | 0.2384 |
| Morocco |  |  | 0.3793 | 0.3645 |
| Pakistan |  |  | 0.1020 | 0.0699 |
| Peru |  |  | 0.6066 | 0.3893 |
| Philippines | -0.5521 | 0.6393 | 1.0249 | 0.4946 |
| Poland |  |  | 0.2287 | 0.2664 |
| Romania |  |  | -0.0605 | 0.4791 |
| Russia |  |  | 2.0476 | 0.5626 |
| S. Africa | -1.6594 | 1.3968 | 0.1044 | 0.4304 |
| S. Arabia | -0.073 | 0.0573 |  |  |
| S. Korea |  |  | 0.5437 | 0.5122 |
| Singapore | 0.5269 | 0.4559 | 0.8214 | 0.5338 |
| Slovakia |  |  | -2.3804 | 1.1016 |
| Taiwan | 0.5754 | 0.4218 | 0.5437 | 0.4397 |
| Thailand | -1.1901 | 0.6349 | 0.2413 | 0.9797 |
| Turkey | 0.0103 | 0.0272 | 0.0059 | 0.0227 |

The table presents estimates of the BF regressionâs slope coefficient, denoted by $\hat{\beta}$. The â Frankel and Poonawala (2010) âcolumn reports $\hat{\beta}$ for the FP sample from a balanced SUR model including developed and developing countries. The âFull Sampleâcolumn reports $\hat{\beta}$ for the full sample from an unbalanced SUR model including developed and developing countries. The $p$-value presented at the bottom is for a Wald test comparing the two sample periods, where critical values come from a Chi. Sq. distribution. The test is conducted by estimating the unbalanced SUR model with intercept and interaction terms for the Frankel and Poonawala (2010) sample period and testing the joint significance of these interaction terms. Full sample estimates are from an unbalanced SUR regression, whereas the Frankel and Poonawala (2010) sample is balanced (at 78 observations) following Frankel and Poonawala (2010). Numbers in bold indicate rejections of the null that $\beta=1$ at the $5 \%$ level, against the one-sided alternative $\beta<1$.

Table 4. Pooled regressions via unbalanced SUR.

|  | Frankel and Poonawala (2010) |  |  | Full Sample |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $p$-Value $(\beta=\mathbf{0})$ | $p$-Value $(\beta=\mathbf{1})$ |  | $p$-Value $(\beta=\mathbf{0})$ | $p$-Value ( $\beta=\mathbf{1})$ |
| $\hat{\beta}$ Developed | -1.2612 | 0.0200 | 0.0000 | -0.0096 | 0.7140 | 0.0000 |
|  | -0.5404 | 0.0289 |  |  | -0.0262 |  |
| Developing | -0.0405 | 0.4750 | 0.0000 | 0.0512 | 0.0000 |  |

The table presents $\beta$ estimates from unbalanced SUR models of pooled developed and pooled developing countries. Both SUR models are unbalanced, even for the FP sample period, since the number of countries are different for the developed and developing countries. For the full sample, the time periods also differ. The $p$-values in the table are for tests of the developed and developing country estimates. The null hypothesis of $\beta=1$ is tested against the one-sided alternative that the bias is negative. The developing country group excludes Indonesia following Frankel and Poonawala.

### 2.2. The Extended Panel

The regression results reported in Tables 1-4, and those in most of the literature more broadly, should be viewed as descriptive at best. This is because the research disregards the problem of structural change.

The problem can already be seen in the results for the extended panel of countries and sample periods. When based on OLS regressions, we find that $\hat{\beta}$ is considerably less negative in the extended samples for every developed country examined (see the last two columns in Table 1). Four of these countries now have positive estimates and the number of countries that are characterized by negative biasedness at the $5 \%$ level drops from 17 to 12 . The average OLS estimate for the developed country group falls in magnitude from -4.19 to -0.69 .

The OLS results for developing countries shows less evidence of structural change (see Table 2, last three columns). Frankel and Poonawala's countries continue to be characterized largely by positive $\hat{\beta} s$ (although the number falls from ten to eight) in the extended sample periods, with an average estimate for the group that is little changed at 0.32 . The added 21 countries are also characterized by largely positive $\hat{\beta}$ S (14 in total) with an average estimate for the added group that is little different than the original group at 0.27 .

The unbalanced SUR results are suggestive of structural change for both groups of countries (see Table 3, the last two columns). Most developed and developing countries witness higher $\hat{\beta}$ s compared with the original sample (seven of nine and nine of 11, respectively). All of the developing countries that remain from the original sample (recall that we drop Hong Kong and Saudi Arabia) are characterized by a positive full-sample $\hat{\beta}$. A Wald test of the null that the original and extended sample results are the same is rejected at the $1 \%$ level (see the bottom of Table 3). ${ }^{21,22}$

Full-sample slope estimates for the added 17 developing countries (recall that we also dropped four developing euro-area countries and Bahrain and added Indonesia) are similar in character to those for Frankel and Poonawala's original group of countries; the $\hat{\beta}$ s are largely positive (13 of 17 and 11 of 11 , respectively). The added countries' $\hat{\beta}$ s are also mostly less than unity at the $5 \%$ level (10 of 17), although this is the case for only five of the original countries. The extended-panel results continue to show a greater number of negative slope estimates for developed countries. However, Frankel and Poonawala's main finding-that developed countries are characterized by greater negative biasedness-is considerably weakened. The results are reported in the last three

[^71]columns of Table 4. They show that both developed and developing countries are still characterized by a negative forward rate bias. However, the difference between the two country groups is much smaller in the extended panel than in the original panel. Indeed, the developing-country slope estimate is no longer significantly less than zero.

Taken as a whole, the original and extended panel results suggest that the BF regression's results depend on the time period examined.

## 3. The Changing Nature of Forward Rate Biasedness

We now investigate systematically the BF regression's instability across developed and developing countries. Other studies examine this instability, but mostly for a sample of either developed or developing countries. Many studies test for one or more exogenously imposed breaks in a sample of developed countries and report strong evidence of structural change. ${ }^{23}$ Researchers have also used ad hoc procedures to identify subperiods in the data that are characterized by high or low volatility of returns (Clarida et al. (2009)) or money growth (Moore and Roche (2012)). They report $\beta$ estimates that are negative during low volatility regimes and positive during high volatility regimes. Moore and Roche (2012) is one of the few studies that examines the instability for both developed and developing countries. They find that $\hat{\beta}$ varies with the volatility regime for both groups of countries.

Several studies employ endogenous structural change tests that allow for multiple break points at unknown dates. Bekaert and Hodrick (1993) and MacDonald and Nagayasu (2015) estimate a two-state Markov-switching model for the largest developed countries. They find that $\hat{\beta}$ varies with the volatility regime. Bai and Mollick (2010) and Baillie and Cho (2014) employ Bai and Perron's (1998, 2003a, 2003b) sequential test procedure, the former for a sample of developing countries and the latter for a sample of developed countries. They find that financial crisis triggers shifts in $\hat{\beta}$ from negative to positive values. ${ }^{24}$ All of these studies find evidence of multiple break points.

In order to examine the instability of the BF regression, we make use of Bai and Perron's (1998, 2003a, 2003b) sequential test approach. ${ }^{25}$ The test procedure has several advantages for our purposes over the Markov-switching approach. Bekaert and Hodrick's (1993) and MacDonald and Nagayasu's (2015) Markov-switching models assume that any structural change involves a switch between only two possible states, thereby assuming that all structural changes are of the same size. This model also assumes that the timing of these changes is governed by a fixed probability distribution. By contrast, the Bai and Perron procedure leaves open the timing and magnitude of the structural changes in the BF regression.

One limitation of Bai and Perron's test is that its sequential procedure tends to stop too early in the search for breaks (Perron (2006)). To address this problem, we combine the sequential procedure with the double maximum tests proposed by Bai and Perron (1998), following the recommendations of Perron (2006) and Bai and Perron (2003b). We first employ the double maximum test. If the test indicates a break, we continue with the sequential procedure to look for additional break points, conditional on the break point found by the double maximum test. ${ }^{26}$ If neither of the two tests indicate a break for a country, we conclude with no breaks (this was the case only for Bahrain). ${ }^{27}$

The estimated break points, in turn, give rise to multiple subperiods with distinct $\beta$ s. The unbalanced SUR model that we estimate below implies a system in which the number of equations equals the number of countries times the number of distinct subperiods. Small subperiods decrease

[^72]the accuracy of our estimates, while increasing the variance-covariance matrix's size and thus the model's computational demands. To address these problems, we restrict the size of a subperiod to twelve months or longer (with exceptions for the first and the last subperiods). As such, we drop any break point that falls within this minimum bound. ${ }^{28,29}$

In carrying out the structural change analysis, we augment the BF regression with lags of both the left- and right-hand side variables via an $\operatorname{ADL}(2,2)$ specification. The dynamic specification accounts for autocorrelated errors, which are typical with persistent variables like the forward premium. ${ }^{30}$

### 3.1. More Frequent Structural Change

Our structural change analysis examines the BF regression's stability for nine developed countries (including the euro area) and 28 developing countries (Estonia, Latvia, Lithuania, and Slovenia are treated as part of the euro area), including Slovakia, whose sample period runs from March 2002 through November 2005 (recall that we dropped Hong Kong, Saudi Arabia, and Bahrain). Table 5a,b report summary results on the frequency of structural change in our panel. Column 2 of the tables shows the number of break points identified for each country. The full set of change dates is reported in Table A1 in the Appendix A. We find that the frequency of structural change in the BF regression is higher for both developed and developing countries than previously reported. The number of break points for both groups of countries varies considerably, in part because of variation in sample sizes. The number of break points ranges from a low of four to a high of 10 for developed countries and from a low of one to a high of 10 for developing countries. The average number of breaks is 8.67 and 5.32 for the developed and developing groups, respectively.

In order to account for differences in sample sizes, we compute the average number of break points per decade for each country (column 3 in the tables). We would have expected that developing countries would be characterized by considerably more structural change, given their greater vulnerability to macroeconomic and political instability. However, the average number of break points per decade for developed countries is not so different than the number for developing countries, at 2.59 and 3.35, respectively. Table 6 reports a difference in means test for these averages (column 2). The test compares the average number of breaks per decade for developed and developing countries via a two-sample $t$-test. Hence, this is a descriptive result. The test is suggestive that the difference is significant, with a $p$-value equal to 0.053 .

[^73]Table 5. Linear individual country BF regressions for developed countries.

| a |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | \# of <br> Breaks | Ave \# <br> Per Decade | $\begin{gathered} \# \text { of } \\ \beta<1 \end{gathered}$ | $\begin{gathered} \# \text { of } \\ \beta>1 \end{gathered}$ | $\begin{gathered} \text { \# of } \\ \beta<1 \text { Signif. } \end{gathered}$ | \# of $\beta>1$ <br> Signif. | $\begin{gathered} \text { \# of } \\ \beta<1 \text { Per } \\ \text { Decade } \end{gathered}$ | $\begin{gathered} \text { \# of } \\ \beta>1 \text { Per } \\ \text { Decade } \end{gathered}$ |
| Australia | 4 | 1.2869 | 5 | 0 | 1 | 0 | 1.6086 | 0 |
| Canada | 10 | 3.2172 | 7 | 4 | 4 | 1 | 2.252 | 1.2868 |
| Euro Area | 10 | 2.5105 | 9 | 2 | 3 | 0 | 2.2594 | 0.5021 |
| Japan | 7 | 2.1762 | 4 | 4 | 2 | 2 | 1.2435 | 1.2435 |
| Norway | 10 | 3.2172 | 5 | 5 | 4 | 1 | 1.6086 | 1.6086 |
| N. Zealand | 10 | 2.4390 | 8 | 3 | 3 | 2 | 1.9512 | 0.7317 |
| Sweden | 10 | 3.2172 | 7 | 4 | 3 | 1 | 2.2520 | 1.2869 |
| Switzerland | 7 | 2.1762 | 5 | 3 | 1 | 0 | 1.5544 | 0.9326 |
| UK | 10 | 3.1088 | 7 | 4 | 5 | 1 | 2.1762 | 1.2435 |
| AVE | 8.6667 | 2.5944 | 6.3333 | 3.2222 | 2.8889 | 0.8889 | 1.8784 | 0.9817 |
|  |  |  | Columns 4 vs. 5 |  | Columns 6 vs. 7 |  | Columns 8 vs. 9 |  |
| $p$-value |  |  | 0.0007 |  | 0.0022 |  | 0.0006 |  |
| b |  |  |  |  |  |  |  |  |
| Developing | \# of | Ave \# Breaks | \# of | \# of | \# of | \# of | \# of | \# of |
|  |  |  |  |  |  | $\beta>1$ | $\beta<1$ Per | $\beta>1$ Per |
| Country | Breaks | Per Decade | $\beta<1$ | $\beta>1$ | $\beta<1$ Signif. | Signif. | Decade | Decade |
| Argentina | 6 | 5.1064 | 6 | 1 | 4 | 0 | 5.1064 | 0.8511 |
| Brazil | 4 | 2.5806 | 3 | 0 | 2 | 0 | 1.9355 | 0.0000 |
| Bulgaria | 5 | 4.2553 | 4 | 2 | 1 | 1 | 3.4043 | 1.7021 |
| Chile | 6 | 5.1064 | 5 | 2 | 1 | 0 | 4.2553 | 1.7021 |
| China | 5 | 4.7999 | 4 | 2 | 3 | 0 | 3.8400 | 1.9200 |
| Colombia | 3 | 1.7822 | 1 | 3 | 1 | 2 | 0.5941 | 1.7822 |
| Czech Rep. | 6 | 3.1579 | 4 | 3 | 1 | 0 | 2.1053 | 1.5789 |
| Hungary | 7 | 3.8532 | 6 | 2 | 4 | 0 | 3.3028 | 1.1009 |
| India | 8 | 4.4037 | 8 | 1 | 3 | 0 | 4.4037 | 0.5505 |
| Indonesia | 1 | 0.5263 | 1 | 1 | 1 | 0 | 0.5263 | 0.5263 |
| Israel | 1 | 0.5714 | 2 | 0 | 1 | 0 | 1.1429 | 0.0000 |
| Kuwait | 5 | 1.9544 | 2 | 4 | 1 | 2 | 0.7818 | 1.5635 |
| Malaysia | 1 | 0.9600 | 2 | 0 | 1 | 0 | 1.9200 | 0.0000 |
| Mexico | 7 | 3.6842 | 6 | 2 | 2 | 1 | 3.1579 | 1.0526 |
| Morocco | 6 | 5.1064 | 4 | 3 | 3 | 0 | 3.4043 | 2.5532 |
| Pakistan | 3 | 1.6981 | 3 | 1 | 3 | 0 | 1.6981 | 0.5660 |
| Peru | 6 | 5.1064 | 5 | 2 | 3 | 0 | 4.2553 | 1.7021 |
| Philippines | 7 | 3.6842 | 5 | 3 | 5 | 1 | 2.6316 | 1.5789 |
| Poland | 7 | 3.6207 | 5 | 3 | 2 | 1 | 2.5862 | 1.5517 |
| Romania | 6 | 5.1064 | 4 | 3 | 2 | 1 | 3.4043 | 2.5532 |
| Russia | 8 | 6.8085 | 6 | 3 | 2 | 2 | 5.1064 | 2.5532 |
| South Africa | 10 | 3.9088 | 8 | 2 | 1 | 0 | 3.1270 | 0.7818 |
| South Korea | 8 | 4.5070 | 7 | 2 | 3 | 1 | 3.9437 | 1.1268 |
| Slovakia | 2 | 1.4458 | 2 | 0 | 2 | 0 | 1.4458 | 0.0000 |
| Singapore | 5 | 1.6086 | 4 | 2 | 0 | 0 | 1.2869 | 0.6434 |
| Taiwan | 6 | 3.1579 | 5 | 2 | 2 | 1 | 2.6316 | 1.0526 |
| Thailand | 6 | 3.2579 | 6 | 1 | 3 | 0 | 3.2579 | 0.5430 |
| Turkey | $4$ | $2.1053$ | 5 | 0 | 5 | 0 | 2.6316 | $0.0000$ |
| Ave. | 5.3214 | 3.3523 | 4.3929 | 1.7857 | 2.2143 | 0.4643 | 2.7817 | 1.1263 |
| $p$-value |  |  | $\begin{gathered} \text { Columns } 4 \text { vs. } 5 \\ 0.0000 \end{gathered}$ |  | $\begin{gathered} \text { Columns } 6 \text { vs. } 7 \\ 0.0000 \end{gathered}$ |  | $\begin{gathered} \text { Columns } 8 \text { vs. } 9 \\ 0.0000 \end{gathered}$ |  |

Columns 2 and 3 present the number of breaks in $\beta$ in total and on average per decade, respectively, using Bai and Perron's $(1998,2003 a, 2003 b)$ sequential approach. Columns $4-9$ present the number of negative and positive biases based on the unbalanced SUR results for developed countries. We lose one subperiod for Norway due to multicollinearity. The reported $p$-values are for two sample $t$-tests and compare the listed columns. The tests should be viewed as descriptive; they test the average number of the respective occurrences and are based on structural break tests that do not account for dependence in the data.

Table 6. Difference in means tests.

|  | Ave \# of Breaks Per Decade | \# of Pos/Neg Biases | (\# of Pos/Neg Biases )/\# of Decades |
| :---: | :---: | :---: | :---: |
| Developed | 2.5944 | 0.5457 | 0.1680 |
| Developing | 3.3523 | 0.5230 | 0.3171 |
| $\boldsymbol{p}$-value | 0.0531 | 0.8891 | 0.0598 |

The table compares the developed and developing countriesâaverage number of breaks per decade, the mean ratio of positive to negative biases, and the mean ratio of positive to negative biases per decade. The reported $p$-values are for two-sample $t$-tests. The tests should be viewed as descriptive; they test the average number of the respective occurrences and are based on structural break tests that do not account for dependence on the data.

## 3.2. $\beta$ Is Not Always Less Than Unity

The break points reported in Table A1 imply multiple subperiods or regimes for which the hypothesis of no structural change cannot be rejected. We use these results and estimate the

BF regression in the distinct regimes for each country in our panel. In order to correct for correlation in the data, we make use of an unbalanced SUR model, which is estimated by the approach described by McDowell (2004). The regression estimates the subperiod $\hat{\beta}$ s for all developed and developing countries as a system, thereby accounting for correlations across countries and making the model temporally nonlinear.

The subperiod estimates give rise to a piecewise linear specification of returns for each country. If the subperiod $\beta$ s were uniformly less than unity, betting against the forward rate would continue to yield predictable profits despite the slope coefficient's instability (assuming $\alpha=0$ ). Bekaert and Hodrick's (1993) and MacDonald and Nagayasu's (2015) Markov-switching results suggest that this is the case; they find a $\beta<1$ for both states of nature.

However, test procedures that leave open the number, timing, and magnitude of the structural changes show that $\beta$ is not always less than unity. The full set of subsample estimates for each country are reported in Table A1. We summarize the results in the last six columns of Table 5a,b. Columns 4 and 5 report the number of regimes for which the estimated bias is found to be less than and greater than unity, respectively. Columns 6 and 7 show the number of estimated biases that are significantly negative and positive (at the $5 \%$ level), respectively, while columns 8 and 9 report the number of negative and positive biases per decade.

We find that $\beta=1$ cannot be rejected in roughly half of the regimes for most countries in both groups. This finding is not surprising given the much shorter subperiod samples. We also find that nearly all countries are characterized by multiple regimes in which $\hat{\beta}<1$ and other regimes in which $\hat{\beta}>1 .{ }^{31}$

The pronounced instability vitiates the view that forward rate biasedness is more negative for developed than developing countries. Moreover, assuming away the instability misses revealing patterns in the biasedness across the two groups of countries. These patterns shed light on the literature's linear regression estimates. They also point towards the importance of risk in driving currency markets.

### 3.3. Two Key Patterns

Two key patterns emerge from the subperiod slope estimates. One pattern is shared by both developed and developing countries: regimes with a negative bias occur more frequently than those with a positive bias. The other pattern reveals an important difference across the two country groups: the size of the subperiod biases, both negative and positive, is much larger for developed countries.

### 3.3.1. Negative-Bias Subperiods Are More Frequent

Table 5a,b, columns 4 and 5, show that the regimes in which $\hat{\beta}<1$ occur roughly twice as often in our sample on average as those in which $\hat{\beta}>1$ for both country groups: 6.33 and 3.22 , respectively, for developed countries and 4.39 and 1.79 , respectively, for developing countries. Difference in means tests (which are reported at the bottom of the tables and again should be treated as descriptive) indicate that the number of negative biases is significantly higher than the number of positive biases for both groups of countries. Moreover, the average of the ratios of positive to negative estimated biases for developed countries is little different than that for developing countries, at 0.55 and 0.52 , respectively. These averages are reported in Table 6, column 3, along with a difference in means test, which suggests a small difference. Again, the tests in Table 6 should be viewed only as descriptive since they do not account for the dependence across countries.

The greater number of negative estimated biases for both groups of countries could be a result of sample sizes. To check this possibility, we examine the number of positive and negative biases

[^74]per decade for each country (columns 8 and 9 in the tables). We find that the scaled figures deliver a similar result: the number of negative biases on average is roughly twice the number of positive biases for developed and developing countries, respectively. We also examine the ratio of positive to negative estimated biases for each country per decade. Table 6 (column 4) reports the average ratio for developed and developing countries, along with a difference in means test. The test is suggestive that the difference is not large.

The higher frequency of negative-bias subperiods is suggestive that carry trade strategies, which bet against the forward rate, are profitable over prolonged stretches of time in both developed and developing countries. Burnside et al. $(2007,2011 b)$, and others report such profitability. However, our structural change results imply that carry trade returns are time dependent and risky. ${ }^{32}$ Time periods that are characterized by a $\hat{\beta}<1$ and carry trade profits are eventually followed by a time period with a $\hat{\beta}>1$ and carry trade losses.

### 3.3.2. Developed Countries Have Larger Biases

Tables 7 and 8 (columns 2 and 3) report the average size of the estimated negative and positive subperiod biases for the two country groups. We find that the negative and positive biases for developed countries- -4.11 and 7.25 , respectively—are roughly 1.5 and 4 times larger on average, respectively, than the negative and positive biases for developing countries- -2.94 and 1.85 , respectively. We see similar patterns when we compute averages of only the significant $\hat{\beta}$ s (columns 4 and 5 in the tables) or consider a weighted average of the $\hat{\beta}$ s using as weights the number of observations in a subperiod (columns 6 and 7 in the tables). Statistical tests reported in Table 9 (columns 2-5) show that the size differences in both the negative and positive biases across developed and developing countries are largely significant. The tests take into account the variance-covariance matrix of the unbalanced nonlinear SUR regression. ${ }^{33}$

Table 7. Average forward rate bias across regimes: developed countries.

| Full Sample | Ave | Ave | Ave Signif | Ave Signif | WAve | WAve |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Developed Country | $(\hat{\beta}-\mathbf{1})<\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})>\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})<\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})>\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})<\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})>\mathbf{0}$ |
| Australia | -2.9914 | NA | NA | NA | -1.7228 | NA |
| Canada | -5.9271 | 6.7986 | -20.58 | 18.6943 | -3.9874 | 4.3390 |
| Euro Area | -3.0035 | 2.0495 | -4.2737 | 1.8425 | -2.861 | 2.0519 |
| Japan | -3.1328 | 22.4420 | -2.4243 | 43.5799 | -2.2822 | 12.4000 |
| Norway | -1.8024 | 4.3062 | NA | 14.1781 | -1.9779 | 3.7675 |
| N. Zealand | -5.2393 | 3.7370 | -5.4428 | 4.7345 | -4.0872 | 3.9194 |
| Sweden | -2.3979 | 5.6082 | -4.8378 | 18.3950 | -2.6548 | 5.3595 |
| Switzerland | -8.4811 | 6.9277 | NA | NA | -3.9068 | 6.9442 |
| UK | -4.0296 | 6.1613 | -8.5512 | 3.5790 | -3.9248 | 5.2884 |
| AVE | -4.1117 | 7.2538 | -7.685 | 15.0005 | -3.0450 | 5.5087 |

The table compares the developed and developing countriesâaverage number of breaks per decade, the mean ratio of positive to negative biases, and the mean ratio of positive to negative biases per decade. The reported $p$-values are for two-sample $t$-tests. The tests should be viewed as descriptive; they test the average number of the respective occurrences and are based on structural break tests that do not account for dependence in the data.

[^75]Table 8. Average forward rate bias across regimes: developed countries.

| Full Sample | Ave | Ave | Ave Signif | Ave Signif | WAve | WAve |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Developing Country | $(\hat{\beta}-\mathbf{1})<\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})<\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})<\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})<\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})<\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})<\mathbf{0}$ |
| Argentina | -0.5385 | 0.1154 | -0.427 | 0.1154 | -0.4789 | 0.1154 |
| Brazil | -0.3912 | NA | -0.5771 | NA | -0.148 | NA |
| Bulgaria | -15.6032 | 8.1199 | NA | NA | -13.6887 | 9.6010 |
| Chile | -9.5817 | 0.4332 | -18.2676 | 0.3136 | -10.0681 | 0.3785 |
| China | -0.8983 | 1.2445 | -0.925 | 0.6309 | -0.8086 | 1.0236 |
| Colombia | -2.014 | 1.9693 | -2.014 | 1.9693 | -2.014 | 1.0956 |
| Czech Rep. | -8.9897 | 3.1573 | -5.9186 | NA | -6.6408 | 3.1830 |
| Hungary | -2.886 | 0.3530 | -2.4584 | 0.0809 | -2.1472 | 0.2909 |
| India | -1.3612 | 0.2974 | -2.2732 | NA | -1.2374 | 0.2974 |
| Indonesia | -0.8434 | 2.8981 | NA | 2.8981 | -0.8434 | 2.8981 |
| Israel | -1.0081 | NA | NA | NA | -1.0689 | NA |
| Kuwait | -1.4426 | 1.8809 | NA | 1.8809 | -1.3513 | 1.3770 |
| Malaysia | -3.1463 | NA | NA | NA | -2.3838 | NA |
| Mexico | -0.9832 | 3.1531 | NA | 5.8058 | -1.0102 | 2.2689 |
| Morocco | -1.2506 | 0.5352 | NA | 0.4733 | -1.2125 | 0.5333 |
| Pakistan | -0.6239 | 0.3060 | -0.4678 | 0.3060 | -0.4954 | 0.3060 |
| Peru | -1.926 | 0.9912 | -3.1029 | 0.5442 | -2.263 | 0.9142 |
| Phippines | -3.0746 | 0.8712 | -3.0746 | 1.5622 | -2.4908 | 0.9241 |
| Poland | -1.8891 | 4.6053 | -0.3618 | 8.3328 | -1.4679 | 4.6682 |
| Romania | -1.6979 | 3.5425 | NA | 9.5916 | -1.6564 | 3.3703 |
| Russia | -1.4876 | 1.4999 | NA | 1.8991 | -1.6152 | 1.7428 |
| S. Africa | -0.7914 | 1.0140 | -0.1457 | 1.8378 | -0.5176 | NA |
| S. Korea | -2.2052 | 1.0901 | -9.541 | 1.9395 | -1.9658 | 1.2115 |
| Slovakia | -4.0233 | NA | -3.8608 | NA | -4.0197 | NA |
| Singapore | -0.3186 | 2.2261 | NA | 1.0443 | -0.3005 | 1.5448 |
| Taiwan | -10.6036 | 1.7152 | -48.4316 | 3.3931 | -5.2055 | 1.5626 |
| Thailand | -1.7939 | 0.4889 | -3.1215 | 0.4889 | -1.687 | 0.4889 |
| Turkey | -0.8314 | NA | -0.7207 | NA | -0.8047 | NA |
| AVE | -2.9359 | 1.8482 | -5.8716 | 2.2554 | -2.4854 | 1.8089 |

The table presents the estimates from the unbalanced SUR model. Ave $((\hat{\beta})-1)$ is the straight average of the sub-period forward rate biases. Ave Signif takes into account only the $5 \%$ significant $\hat{\beta}$ estimates. WAve denotes a weighted average using the number of observations in each subperiod for weights. NA denotes not applicable due to lack of observations.

Table 9. Difference in means tests.

|  | Ave | Ave | Ave <br> Signif | Ave <br> Signif | Ave <br> 1990s-2016 | Ave <br> Ave Signif. <br> 1990s-2016 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  | Ave Signif. |  |  |  |
|  | $(\hat{\beta}-\mathbf{1})<\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})>\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})<\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})>\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})<\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})>\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})<\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})>\mathbf{0}$ |
| Developed | -4.1117 | 7.2538 | -7.685 | 15.0005 | -5.2049 | 7.5580 | -8.7729 | 17.7866 |
| Developing | -2.9359 | 1.8482 | -5.8716 | 2.2554 | -2.9347 | 1.8482 | -5.8716 | 2.2554 |
| $p$-value | 0.1581 | 0.0006 | 0.0002 | 0.0000 | 0.0610 | 0.0000 | 0.0540 | 0.0001 |

The table presents the averages of the negative and positive biases for developed and developing countries. Ave is the straight average of the sub-period forward rate biases. Ave Signif takes into account only the $5 \%$ significant $\hat{\beta}$ estimates. The reported $p$-values are for Wald tests comparing the biases of developed and developing countries, where the critical values come from the Chi Sq. distribution.

### 3.3.3. Origins of Linear Regression Results

The two key patterns show the origins of the literature's linear regression findings. The negative biasedness found for both developed and developing countries arises from the greater frequency of regimes for which $\hat{\beta}<1$. The larger negative bias that Frankel and Poonawala (2010) and others report for developed countries stems from the much larger subperiod biases for this group.

The results for the developed countries stem in part from the behavior of the 1980s, which were characterized by large swings in dollar exchange rates. The developing-country samples miss this
period, which may underpin the larger biases for developed countries. ${ }^{34}$ In order to check this possibility, we analyze the biases of the developed countries for truncated samples that start in the 1990s. ${ }^{35}$

Table 10 reports these results. We find that the size of the estimated biases for developed countries is, for the most part, even greater than when the 1980s are omitted from the sample. They continue to show that the size of both negative and positive biases is larger for developed than developing countries. The last four columns of Table 9 show that the differences continue to be significant for this time period.

Table 10. Average forward rate bias across regimes: developed countries.

| 1990s-2016 | Ave | Ave | Ave Signif | Ave Signif | WAve | WAve |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Developed Country | $(\hat{\beta}-\mathbf{1})<\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})>\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})<\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})>\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})<\mathbf{0}$ | $(\hat{\beta}-\mathbf{1})>\mathbf{0}$ |
| Australia | -3.4122 | NA | NA | NA | -1.1091 | NA |
| Canada | -6.7085 | 6.7986 | -20.58 | 18.6943 | -3.8371 | 4.3390 |
| Euro Area | -4.2955 | 2.2566 | -4.8451 | NA | -2.3744 | 1.1410 |
| Japan | -3.881 | 22.4420 | -2.4243 | 43.5799 | -1.9602 | 12.4000 |
| Norway | -1.7755 | 4.3062 | NA | 14.1781 | -1.5509 | 3.7675 |
| N. Zealand | -8.6162 | 5.6367 | -5.4428 | 5.6367 | -3.2916 | 2.0733 |
| Sweden | -3.0723 | 5.6082 | -6.216 | 18.3950 | -2.1244 | 5.3595 |
| Switzerland | -13.1124 | 6.9277 | NA | NA | -3.5167 | 6.9442 |
| UK | -3.3261 | 6.1613 | NA | 3.5790 | -2.5905 | 5.2884 |
| AVE | -5.3556 | 7.5172 | -7.9016 | 17.3438 | -2.4839 | 5.1641 |

The table presents the estimates from the unbalanced SUR model. Sample periods begin after the first break point in the 1990s for each country. Ave $((\hat{\beta})-1)$ is the straight average of the sub-period forward rate biases. Ave Signif takes into account only the $5 \%$ significant $\hat{\beta}$ estimates. WAve denotes a weighted average using the number of observations in each subperiod as weights. NA denotes not applicable due to lack of observations.

## 4. Unpredictability and Imperfect Knowledge

The pronounced instability of forward rate biasedness implies that currency returns are not predictable on the basis of the linear BF model. However, currency returns may nonetheless be predictable. The question turns on whether the instability can be modeled ex ante with a probability rule. Goldberg et al. (Forthcoming) examine this question for six developed-country markets. They consider the out-of-sample predictive performance of Markov-switching and other nonlinear regression models. They find that the nonlinear models have little or no predictive power. ${ }^{36}$

This finding is supportive of the main premise of imperfect knowledge economics: the process underpinning economic outcomes undergoes change at points in time and in ways that cannot be characterized ex ante with the same probability rule at all points in time. Frydman and Goldberg (2013a, 2013b); Frydman et al. (2015) argue that this Knightian uncertainty arises because structural change in financial markets and the broader economy is triggered in part by historical developments that are to some extent novel. Examples include the appointment of a new central bank governor or Treasury Secretary, shifts in exchange rate policy, German reunification, and financial crises. The novelty of these events implies that they are to some extent non-repetitive and that their impact on returns "deal[s] with situations which are far too unique...[to rely solely on] statistical tabulations" Knight (1921), p. 198.

Consequently, the structural shifts that they trigger are unlikely to be characterized by a stable probability rule.

[^76]The corpus of Johansen and Juselius's empirical work on currency markets provides considerable evidence of a connection between structural change and novel historical developments. In nearly all of their CVAR studies, they have had to include equilibrium mean shifts, broken trends, and a series of various dummies to account for the impact of major policy changes and other novel historical developments. Juselius (2017b) is a case in point. The study estimates a CVAR for the German mark-U.S. dollar exchange rate. It finds that a well specified unrestricted model requires: (1) a broken trend and step dummy in January 1991 to control for German reunification; (2) an impulse dummy to account for three new German excise taxes to pay for reunification; and (3) intervention dummies to account for large shocks to U.S. goods and bond prices in the aftermath of the 1985 Plaza Accord.

There is also considerable evidence in the broader literature that structural change in financial markets is triggered by novel historical developments. ${ }^{37}$ These developments are not mere repetitions of events in the past. As such, no probabilistic rule that was estimated on ex post data are likely to enable one to predict their timing or character, let alone their quantitative impact on the process underpinning asset returns.

Our structural change results provide additional evidence of a connection between structural change and novel historical developments. The full set of break points for developed and developing countries (see Table A1) show that breaks are proximate to major historical developments. For example, one or more break points that are proximate to the 2008 global financial crisis are found for nearly all countries. In many cases, the instability involved a change in the sign of the forward rate bias. Not surprisingly, Baillie and Cho (2014) and Daniel et al. (2017) find that carry trade strategies produced large losses after 2008. Few economists or market participants predicted the financial crisis, let alone its impact on currency markets and forward rate bias.

## 5. Developed Countries Are More Risky Not Less Rational

The BF regression's instability implies that conclusions based on the linear model about the rationality or irrationality of currency markets in general, or developed countries' relative irrationality in particular, are unfounded. Moreover, the structural change results in Tables 5-9 point instead to the importance of risk in driving currency returns and in understanding forward rate biasedness across developed and developing countries.

A common measure of risk in financial markets is the volatility of returns. However, volatility measures suffer from the problem of structural change: the basic properties of market data are not the same during crisis and noncrisis periods (e.g., see Jorion (2009)). Indeed, market volatility tends to fall during large price upswings away from benchmark values. However, it is precisely during these periods that risk rises, as markets become more vulnerable to sudden structural change that is accompanied by large and sustained price reversals. This behavior lends support to IKE's alternative risk premium model, which implies that growing departures from benchmark values lead to greater risk for investors who bet on even larger departures from benchmark values. The findings of Juselius and Assenmacher (2017), Juselius and Stillwagon (2018), Cavusoglu et al. (2020), and others provide strong support for this alternative measure of risk.

This research suggests that the frequency and magnitude of the instability in the BF regression provide useful additional measures of market risk. ${ }^{38}$ In terms of the forward rate puzzle, the carry trade's ability to generate profits or losses depends on the sign and size of the bias. A negative bias implies profits on average, whereas a positive bias implies losses. In addition, the larger the size of the

[^77]bias, the greater are the positive or negative returns. Consequently, a greater frequency or magnitude of structural change in the BF regression implies greater risk in betting against (or with) the forward rate.

The instability results reported in Table 5a,b show that currency markets and carry trade strategies are riskier than what is suggested by the literature's linear regression results. In terms of the prevalence of structural change, and the frequency with which negative- and positive-bias regimes occur, developed and developing countries are comparably risky.

However, as we saw in Tables 7-10, the size of both the negative and positive biases is larger for developed countries than developing countries. This finding implies greater profits to carry trade strategies in developed countries during regimes with a negative bias, but also greater losses during regimes with a positive bias. It also suggests that the magnitude of the structural changes that occur in developed-country markets is on average much greater than in developing-country markets.

We find that this is the case. The size difference is easy to spot in Figure 2, which provides frequency distributions of the absolute value of the changes in $\hat{\beta}$ that occurred across the distinct subperiods for each country, one for the developed countries (the solid blue line), and the other for developing countries (the dashed red line). We find that the mean of the structural changes for the group of developed countries (5.08) is roughly twice the mean change for the group of developing countries (2.59). A descriptive difference in means test is suggestive that the difference is large. We also employ an ks-test, which is suggestive that the distributions for the two country groups are different.


Figure 2. Absolute changes in $\beta$ across subperiods.
The results imply that betting against (with) the forward rate in developed-country markets delivers higher profits than in developing-country markets during subperiods in which the forward
rate bias is negative (positive). However, they also imply that speculation in the major markets is much more risky relative to developing-country markets than previously thought.

## 6. Conclusions

The structural change results reported in this paper indicate that the literature's linear-regression findings lead to spurious conclusions about the importance of rationality and risk in developedand developing-country currency markets. The BF regression's instability and lack of predictive power imply that the widespread claims of currency markets' irrationality are premature at best. The connection between structural change and novel historical developments suggests that imperfect knowledge and not irrationally is key to understanding these markets. The instability findings also indicate that developed-country markets are riskier than those in developing countries.

An open question is whether any of the risk factors considered in the literature can explain (1) why subperiods with a negative forward rate bias occur nearly twice as often as subperiods with a positive bias; and (2) why the subperiod biases (both negative and positive) are on average much larger for developed countries. The findings of Juselius (1992, 1995, 2014, 2017a, 2017b); Johansen and Juselius (1992); Juselius and MacDonald (2004); Juselius and Assenmacher (2017); Johansen et al. (2010) and Juselius and Stillwagon (2018) suggest that the IKE gap measure of risk is a good place to start.

Author Contributions: Conceptualization and methodology, M.D.G. and O.K.; empirical analysis, O.K. and D.O.; writing, M.D.G., O.K., D.O.; All authors have read and agreed to the published version of the manuscript.
Funding: This research received funding from the Institute for New Economic Thinking and the University of New Hampshire.
Conflicts of Interest: The authors declare no conflict of interest.

## Appendix A

Table A1. Subperiod coefficients and corresponding break dates.

| Country/Subperiod | Total \# of Observations | Coefficient | Std. Error | Break Dates | Number of Months per Subperiod |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Developed |  |  |  |  |  |
| Australia | 373 | -0.3082 | 0.5984 | 1999-06 | 175 |
|  |  | -8.6771 | 13.7382 | 2001-02 | 20 |
|  |  | -0.6493 | 1.1689 | 2008-06 | 88 |
|  |  | -0.7285 | 2.8214 | 2010-1 | 19 |
|  |  | 0.4062 | 0.9101 |  | 71 |
| Canada | 373 | -0.2389 | 0.9779 | 1987-04 | 29 |
|  |  | -1.4654 | 1.0122 | 1989-11 | 31 |
|  |  | 0.2372 | 0.6406 | 1993-08 | 45 |
|  |  | -2.6832 | 2.1780 | 1995-04 | 20 |
|  |  | -1.2089 | 1.5584 | 2001-10 | 78 |
|  |  | -9.5507 | 5.0584 | 2003-04 | 18 |
|  |  | 1.0208 | 3.9075 | 2006-02 | 34 |
|  |  | 7.5901 | 4.6161 | 2007-09 | 19 |
|  |  | -19.5800 | 9.8565 | 2009-03 | 18 |
|  |  | 2.8892 | 2.4325 | 2014-06 | 63 |
|  |  | 19.6943 | 7.9364 |  | 18 |
| Euro Area | 478 | 0.4528 | 0.7482 | 1978-11 | 33 |
|  |  | 0.0321 | 0.9964 | 1981-09 | 34 |
|  |  | -2.1309 | 0.9799 | 1984-05 | 32 |
|  |  | 2.8425 | 1.1327 | 1988-1 | 44 |
|  |  | 0.0923 | 0.8849 | 1991-07 | 42 |
|  |  | 3.2566 | 4.2712 | 1995-04 | 45 |
|  |  | -2.4436 | 1.1624 | 2000-12 | 68 |
|  |  | -5.2466 | 2.1608 | 2005-1 | 49 |
|  |  | 0.4328 | 1.5662 | 2008-12 | 47 |
|  |  | -1.8837 | 5.8379 | 2014-03 | 63 |
|  |  | -7.3366 | 8.4644 |  | 21 |

Table A1. Cont.

| Country/Subperiod | Total \# of Observations | Coefficient | Std. Error | Break Dates | Number of Months per Subperiod |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Developed |  |  |  |  |  |
| Japan | 386 | 0.1118 | 1.0256 | 1990-09 | 83 |
|  |  | 1.8287 | 2.5877 | 1995-03 | 54 |
|  |  | -1.9918 | 1.1511 | 1998-07 | 40 |
|  |  | 2.7793 | 1.8288 | 2000-02 | 19 |
|  |  | -1.4243 | 0.6852 | 2007-05 | 87 |
|  |  | 5.6131 | 2.6862 | 2012-10 | 65 |
|  |  | 83.5468 | 30.6330 | 2014-05 | 19 |
|  |  | -5.2268 | 9.2788 |  | 19 |
| New Zealand | 492 | -2.2846 | 1.3259 | 1978-02 | 39 |
|  |  | 4.8323 | 1.8153 | 1980-09 | 31 |
|  |  | -1.0799 | 1.8461 | 1984-10 | 49 |
|  |  | 2.7421 | 6.3360 | 1986-10 | 24 |
|  |  | -0.6085 | 2.6035 | 1989-07 | 33 |
|  |  | 0.5230 | 0.6438 | 1996-05 | 82 |
|  |  | -3.7483 | 2.4294 | 1998-05 | 24 |
|  |  | -4.4428 | 1.6941 | 2006-06 | 97 |
|  |  | 6.6367 | 3.2458 | 2009-02 | 32 |
|  |  | -20.5526 | $24.2922$ | 2011-03 | 25 |
|  |  | $-1.7211$ | $5.4291$ |  | 56 |
| Norway | 373 | -0.9166 | 1.0722 | 1986-11 | 24 |
|  |  | -0.7687 | 0.6134 | 1989-12 | 37 |
|  |  | 0.0000 | NA | 1991-06 | 18 |
|  |  | 0.5360 | 1.0065 | 1993-04 | 22 |
|  |  | 2.7532 | 2.9482 | 1995-02 | 22 |
|  |  | -1.7781 | 1.5888 | 2001-06 | 76 |
|  |  | -1.0843 | 0.8722 | 2008-06 | 84 |
|  |  | 5.1275 | 3.4849 | 2009-12 | 18 |
|  |  | 1.6363 | 2.7222 | 2011-11 | 23 |
|  |  | 1.8358 | 2.6974 | 2014-05 | 30 |
|  |  | 15.1781 | 6.6951 |  | 19 |
| Sweden | 373 |  |  |  |  |
|  |  | $0.7971$ | 2.1823 | 1989-08 | 21 |
|  |  | 0.1666 | 0.9902 | 1991-07 | 23 |
|  |  | 2.5192 | 1.6961 | 1993-1 | 18 |
|  |  | -1.0084 | 1.2543 | 1998-12 | 71 |
|  |  | -2.9010 | 1.8458 | 2001-05 | 29 |
|  |  | -5.2160 | 1.5676 | 2005-10 | 53 |
|  |  | $2.6345$ | 1.4783 | 2008-06 | 32 |
|  |  | 19.3950 | 5.7027 | 2010-04 | 22 |
|  |  | 0.8363 | 2.2597 | 2014-02 | 46 |
|  |  | 1.8842 | 7.1255 |  | 22 |
| Switzerland | 386 | -0.4233 | 0.9348 | 1986-08 | 34 |
|  |  | -0.6516 | 1.2321 | 1990-07 | 47 |
|  |  | 3.2036 | 3.0870 | 1992-03 | 20 |
|  |  | -0.8667 | 0.7235 | 2007-05 | 182 |
|  |  | 4.2475 | 4.1783 | 2009-03 | 22 |
|  |  | 16.3319 | 12.9444 | 2010-12 | 21 |
|  |  | -32.6304 | 27.5562 | 2012-07 | 19 |
|  |  | -2.8336 | 6.4644 |  | 41 |
| UK | 386 |  |  | 1985-06 |  |
|  |  | -1.8967 | 1.3367 | 1987-05 | 23 |
|  |  | -2.8908 | 2.3169 | 1988-12 | 19 |
|  |  | -1.8905 | 1.2130 | 1990-09 | 21 |
|  |  | 1.1869 | 1.2998 | 1992-07 | 22 |
|  |  | 4.5790 | 1.5620 | 1994-06 | 23 |
|  |  | -2.2244 | 1.3260 | 2003-04 | 106 |
|  |  | -0.8953 | $2.4753$ | 2004-11 | 19 |
|  |  | 4.4470 | 2.3033 | 2008-06 | 43 |
|  |  | 18.4322 | 10.6962 | 2010-1 | 19 |
|  |  | -3.8585 | 6.7086 |  | 71 |
| Developing |  |  |  |  |  |
| Argentina | 141 |  | $0.5646$ |  |  |
|  |  | 0.7673 | 0.3107 | 2007-12 | 37 |
|  |  | 0.2887 | 0.0363 | 2008-12 | 12 |
|  |  | 0.4860 | 0.1059 | 2012-06 | 42 |
|  |  | 1.1154 | 0.2744 | 2013-12 | 18 |
|  |  | 0.3883 | 0.0672 | 2015-06 | 18 |
|  |  | 0.9347 | 0.0776 |  | 6 |

Table A1. Cont.

| Country/Subperiod | Total \# of Observations | Coefficient | Std. Error | Break Dates | Number of Months per Subperiod |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Developed |  |  |  |  |  |
| Brazil |  |  |  | 30 |  |
|  |  |  | 0.9088 | 0.0403 | $2002-12$ |

Table A1. Cont.

| Country/Subperiod | Total \# of Observations | Coefficient | Std. Error | Break Dates | Number of Months per Subperiod |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Developed |  |  |  |  |  |
| Mexico | 228 | 0.6822 | 0.4218 | 1998-08 | 20 |
|  |  | -0.1806 | 0.2565 | 2001-02 | 30 |
|  |  | 0.3187 | 0.7664 | 2003-02 | 24 |
|  |  | -0.1820 | 0.4709 | 2006-05 | 39 |
|  |  | 1.5004 | 1.0725 | 2009-1 | 32 |
|  |  | -1.0894 | 1.3735 | 2011-06 | 29 |
|  |  | 0.5518 | 1.1641 | 2014-08 | 38 |
|  |  | 6.8058 | 2.3260 |  | 16 |
| Morocco | 141 | -0.1280 | 1.0689 | 2005-02 | 11 |
|  |  | 0.1180 | 0.4491 | 2008-05 | 39 |
|  |  | -0.1794 | 0.5143 | 2009-11 | 18 |
|  |  | 1.8136 | 0.8844 | 2011-05 | 18 |
|  |  | 1.6589 | 0.8872 | 2012-07 | 14 |
|  |  | $-0.8132$ | $0.6044$ |  | $24$ |
|  |  | 1.1329 | 0.5115 |  | 17 |
| Pakistan | 212 | 0.0638 | 0.0573 | 1999-04 | 12 |
|  |  | 1.3060 | 0.4875 | 2000-07 | 15 |
|  |  | 0.4743 | 0.2151 | 2008-02 | 91 |
|  |  | 0.5902 | 0.1446 |  | 94 |
| Peru | 141 | 0.0831 | 0.4122 | 2005-07 | 16 |
|  |  | -2.2890 | 1.7241 | 2007-08 | 25 |
|  |  | 2.4383 | 1.5624 | 2008-08 | 12 |
|  |  | -0.8687 | 1.5976 | 2009-09 | $13$ |
|  |  | -2.1029 | 0.5409 | 2013 | 40 |
|  |  | 0.5476 | 0.3679 | 2014-07 | 18 |
|  |  |  |  |  |  |
| Philippines | 228 | 2.5033 | 0.8389 | 1998-08 | 20 |
|  |  | -0.9631 | 0.8221 | 2000-04 | 20 |
|  |  | 1.4522 | 0.8622 | 2001-06 | 14 |
|  |  | 0.3668 | 0.3465 | 2004-11 | 41 |
|  |  | -2.0437 | 0.9682 | 2006-05 | 18 |
|  |  | -6.1869 | 3.1264 | 2008-1 | 20 |
|  |  | $1.6098$ | 1.1308 | 2009-07 | $18$ |
|  |  | $-1.1853$ | $0.6464$ |  | $77$ |
| Poland | 232 | 0.6382 | 0.2803 | 2000-09 | 49 |
|  |  | -0.5078 | 0.5407 | 2004-04 | 43 |
|  |  |  | 1.9701 | 2005-06 | 14 |
|  |  | -1.2795 | 3.8595 | 2008-06 | $36$ |
|  |  | 4.6760 | 2.9489 | 2010-11 | 29 |
|  |  | 2.8072 | 3.2026 | 2011-12 | 13 |
|  |  | $0.1432$ | $1.4978$ | 2014-06 | $30$ |
|  |  | 9.3328 | 3.6071 |  | 18 |
| Romania | 141 | -1.0197 | 0.5550 | 2005 | 10 |
|  |  | -0.4703 | 1.7485 | 2007-10 | 33 |
|  |  | 1.6233 | 1.2055 | 2009-02 | 16 |
|  |  | -0.1026 | 0.7470 | 2010-08 | 18 |
|  |  | 1.4124 | 1.5923 | 2012-07 | 23 |
|  |  | -1.1989 | 1.2713 | 2014-06 | 23 |
|  |  | 10.5916 | 3.2499 |  | 18 |
| Russia | 141 |  |  |  |  |
|  |  | $-1.2353$ | 1.1216 | 2006-10 | 16 |
|  |  | -2.1772 | 1.7489 | 2008-06 | 20 |
|  |  | $2.0998$ | 0.5678 | 2009-07 | 13 |
|  |  | -1.2113 | 1.9413 | 2010-08 | 13 |
|  |  | 0.2709 | 3.1239 | 2011-09 | 13 |
|  |  | 0.6228 | 1.4569 | 2012-10 | 13 |
|  |  | $3.6984$ | 1.4420 | 2014-10 | $24$ |
|  |  | 1.7015 | 1.5434 |  | 14 |
| South Africa | 307 | 0.8543 | 0.2783 | 1995-12 | 67 |
|  |  | 0.8414 | 0.7293 | 1997-03 | 15 |
|  |  | 2.8378 | 1.2608 | 1998-06 | 15 |
|  |  | 0.8760 | 0.7424 | 2001-08 | 38 |
|  |  | -0.8895 | 1.2953 | 2003-04 | 20 |
|  |  | -2.1292 | 1.5756 | 2004-11 | 19 |
|  |  | 1.1902 | 3.2008 | 2006-04 | 17 |
|  |  | 0.7352 | 2.2410 | 2007-11 | 19 |
|  |  | 0.7370 | 1.8149 | 2009-02 | 15 |
|  |  | $0.6433$ | $0.7325$ | 2013-08 | $54$ |
|  |  | 0.0000 | NA |  | 28 |

Table A1. Cont.

| Country/Subperiod | Total \# of Observations | Coefficient | Std. Error | Break Dates | Number of Months per Subperiod |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Developed |  |  |  |  |  |
| South Korea | 213 | 1.2408 | 1.0681 | 1999-03 | 12 |
|  |  | -8.5410 | 2.2965 | 2000-10 | 19 |
|  |  | -0.3671 | 0.5255 | 2004-09 | 47 |
|  |  | -0.4782 | 0.8521 | 2006-05 | 20 |
|  |  | 0.8593 | 0.4506 | 2008-1 | 20 |
|  |  | 0.6694 | 1.0405 | 2009-12 | 23 |
|  |  | 2.9395 | 0.4449 | 2011-04 | 16 |
|  |  | $0.4545$ | $2.2550$ | 2013-04 | 24 |
|  |  | $-1.0332$ | $1.9164$ |  | 32 |
| Slovakia | 45 | -2.8608 | 1.1622 | 2004-1 | 23 |
|  |  | -3.1858 | 2.0160 | 2005-02 | 22 |
|  |  | 0 | NA |  |  |
| Singapore | 373 | 0.5831 | 0.3100 | 1995-05 | 126 |
|  |  | 0.7140 | 1.3990 | 1997-12 | 31 |
|  |  | 0.9378 | 0.8768 | 2005-09 | 93 |
|  |  | 0.4908 | 0.6365 | 2008-11 | 38 |
|  |  | 2.0443 | 0.8051 | 2014-06 | 67 |
|  |  | 4.4078 | 3.8367 |  | 18 |
| Taiwan | 228 | 4.3931 | 1.4556 | 1998-08 | 20 |
|  |  | 1.0372 | 0.9006 | 2000-08 | 24 |
|  |  | 0.5288 | 0.9851 | 2002-06 | 22 |
|  |  | 0.8003 | 0.4580 | 2008-02 | 68 |
|  |  | -0.3337 | 0.6425 | 2010-12 | 34 |
|  |  | $-1.5819$ | $2.7336$ | 2014-08 | $44$ |
|  |  | -47.4316 | 15.4424 |  | 16 |
| Thailand | 221 |  |  |  | 17 |
|  |  | $-1.1262$ | 1.1897 | 2001-06 | 30 |
|  |  | -3.2605 | 2.2530 | 2002-06 | 12 |
|  |  | $-0.9522$ | $0.4665$ | 2007-06 | 60 |
|  |  | $0.8642$ | 1.1569 | 2008-10 | 16 |
|  |  | -0.9202 | 0.4696 | 2013-03 | 53 |
|  |  | 0.6895 | 0.6123 |  | 33 |
| Turkey | 228 | $0.7387$ | $0.0382$ |  |  |
|  |  | $-0.0219$ | 0.0426 | 2002-1 | 12 |
|  |  | 0.0671 | 0.0286 | 2010-1 | 96 |
|  |  | 0.0322 | 0.0124 | 2014-1 | 48 |
|  |  | 0.0269 | 0.0244 |  | 23 |

Values in bold denote negative and significant biases (at a 0.05 level).

## References

Ahmad, Rubi, Ghon Rhee, and Yuen Meng Wong. 2012. Foreign exchange market efficiency under recent crises: Asia-pacific focus. Journal of International Money and Finance 31: 1574-92. [CrossRef]
Ang, Andrew, and Alan Timmermann. 2012. Regime changes and financial markets. Annual Review of Financial Economics 4: 313-37. [CrossRef]
Ang, Andrew, Joseph Chen, and Yuhang Xing. 2012. Downside risk. Review of Financial Studies 31: 1574-92.
Bai, Shuming, and Andre Varella Mollick. 2010. Currency crisis and the forward discount bias: Evidence from emerging economies under breaks. Journal of International Financial Markets, Institutions and Money 20: 556-74. [CrossRef]
Bai, Jushan, and Pierre Perron. 1998. Estimating and testing linear models with multiple structural changes. Econometrica 66: 47-78. [CrossRef]
Bai, Jushan, and Pierre Perron. 2003a. Critical values for multiple structural change tests. The Econometrics Journal 6: 72-78. [CrossRef]
Bai, Jushan, and Pierre Perron. 2003b. Computation and analysis of multiple structural change models. Journal of Applied Econometrics 18: 1-22. [CrossRef]
Baillie, Richard T., and Tim Bollerslev. 2000. The forward premium anomaly is not as bad as you think. Journal of International Money and Finance 19: 471-88. [CrossRef]
Baillie, Richard T., and Dooyeon Cho. 2014. When carry trades in currency markets are not profitable. Review of Development Economics 18: 794-803. [CrossRef]

Bansal, Ravi, and Magnus Dahlquist. 2000. The forward premium puzzle: different tales from developed and emerging economies. Journal of International Economics 51: 115-44. [CrossRef]
Bansal, Ravi. 1997. An exploration of the forward premium puzzle in currency markets. Review of Financial Studies 10: 369-403. [CrossRef]
Beckmann, Joscha, Ansgar Belke, and Michael Kühl. 2006. The dollar-euro exchange rate and macroeconomic fundamentals: A time-varying coefficient approach. Review of World Economics 147: 11-40. [CrossRef]
Bekaert, Geert, and Robert J. Hodrick. 1993. On biases in the measurement of foreign exchange risk premiums. Journal of International Money and Finance 12: 115-38. [CrossRef]
Bilson, John F. O. 1981. The speculative efficiency hypothesis. Journal of Business 54: 435-51. [CrossRef]
Brunnermeier, Markus K., Stefan Nagel, and Lasse H. Pedersen. 2008. Carry trades and currency crashes. NBER Macroeconomics Annual 23: 313-47. [CrossRef]
Burnside, Craig, Martin S. Eichenbaum, and Sergio Rebelo. 2009. Understanding the forward premium puzzle: A microstructure approach. American Economic Journal: Macroeconomics 2: 127-54. [CrossRef]
Burnside, Craig, Han Bing Hirshleifer David, and Tracy Yue Wang. 2011a. Investor overconfidence and the forward premium puzzle. Review of Economic Studies 78: 523-58. [CrossRef]
Burnside, Craig, Martin S. Eichenbaum, and Sergio Rebelo. 2011b. Carry trade and momentum in currency markets. NBER Working Paper 16942.
Burnside, Craig, Martin S. Eichenbaum, and Sergio Rebelo. 2007. The returns to currency speculation in emerging markets. American Economic Review 97: 333-38. [CrossRef]
Cavusoglu, Nevin, Michael D. Goldberg, and Josh Stillwagon. 2020. Currency Returns and Downside risk: Debt, Volatility, and the Gap from Benchmark Values. Working Paper. Available online: https:/ / www.ineteconomics. org/research/research-papers/new-evidence-on-the-portfolio-balance-approach-to-currency-returns (accessed on 30 October 2020).
Chinn, Menzie D. 2006. The (partial) rehabilitation of interest rate parity in the floating rate era: Longer horizons, alternative expectations, and emerging markets. Journal of International Money and Finance 25: 7-21. [CrossRef]
Clarida, Richard, Josh Davis, and Niels Pedersen. 2009. Currency carry trade regimes: Beyond the fama regression. Journal of International Money and Finance 28: 1373-89. [CrossRef]
Clements, Michael P., and David F. Hendry. 1999. Forecasting Non- Stationary Economic Time Series. Cambridge: The MIT Press.
Daniel, Kent, Robert J. Hodrick, and Zhongjin Lu. 2017. The Carry Trade: Risks and Drawdowns. Critical Finance Review 6: 211-62. [CrossRef]
Deng, Ai, and Pierre Perron. 2008. A non-local perspective on the power properties of the cusum and cusum of squares tests for structural change. Journal of Econometrics 42: 212-40. [CrossRef]
Engel, Charles A. 1996. The forward discount anomaly and the risk premium: A survey of recent evidence. Journal of Empirical Finance 3: 123-91. [CrossRef]
Engel, Charles A. 2014. Exchange rates and interest parity. In Handbook of International Economics. Edited by Gita Gopinath, Elhanan Helpman and Kenneth Rogoff. Amsterdam: North-Holland, vol. IV, pp. 453-522.
Fama, Eugene F. 1984. Forward and spot exchange rates. Journal of Monetary Economics 14: 319-38. [CrossRef]
Flood, Robert P., and Andrew K. Rose. 2002. Uncovered Interest Parity in Crisis. International Monetary Fund Staff Papers 49: 252-66.
Frankel, Jeffrey, and Jumana Poonawala. 2010. The forward market in emerging currencies: Less biased than in major currencies. Journal of International Money and Finance 29: 585-98. [CrossRef]
Froot, Kenneth A., and Richard H. Thaler. 1990. Anomalies: Foreign exchange. Journal of Economic Perspectives 4: 179-92. [CrossRef]
Frydman, Roman, and Michael D. Goldberg. 2007. Imperfect Knowledge Economics: Exchange Rates and Risk. Princeton: Princeton University Press.
Frydman, Roman, and Michael D. Goldberg. 2011. Beyond Mechanical Markets: Asset Price Swings, Risk, and the Role of the State. Princeton: Princeton University Press.
Frydman, Roman, and Michael D. Goldberg. 2013a. Opening models of asset prices and risk to non-routine change. In Rethinking Expectations: The Way Forward for Macroeconomics. Edited by Roman Frydman and Edmund S. Phelps. Princeton: Princeton University Press, pp. 207-50.

Frydman, Roman, and Michael D. Goldberg. 2013b. The imperfect knowledge imperative in modern macroeconomics and finance theory. In Rethinking Expectations: The Way Forward for Macroeconomics. Edited by Roman Frydman and Edmund S. Phelps. Princeton: Princeton University Press, pp. 130-65.
Frydman, Roman, Michael D. Goldberg, and Nicholas Mangee. 2015. Knightian uncertainty and stock-price movements: Why the reh present-value model failed empirically? Economics: The Open-Access, Open-Assessment E-Journal 9: 1-50. [CrossRef]
Goldberg, Michael D., and Roman Frydman. 1996a. Imperfect knowledge and behavior in the foreign exchange market. Economic Journal 106: 869-93. [CrossRef]
Goldberg, Michael D., and Roman Frydman. 1996b. Empirical exchange rate models and shifts in the co-integrating vector. Journal of Structural Change and Economic Dynamics 7: 55-78. [CrossRef]
Goldberg, Michael, Olesia Kozlova, and Peter Sullivan. Forthcoming. Are currency returns really predictable?: Novel structural change and the forward rate anomaly. SSRN.
Gourinchas, Pierre-Olivier, and Aaron Tornell. 2004. Exchange rate puzzles and distorted beliefs. Journal of International Economics 64: 303-33. [CrossRef]
Goyal, Amit, and Ivo Welch. 2008. A comprehensive look at the empirical performance of equity premium prediction. The Review of Financial Studies 21: 1455-508.
Hoover, Kevin D., Katarina Juselius, and Søren Johansen. 2008. Allowing the data to speak freely: The macroeconometrics of the cointegrated vector autoregression. American Economic Review 98: 251-55. [CrossRef]
Ito, Hiro, and Menzie Chinn. 2007. Price-based measurement of financial globalization: A cross-country study of interest rate parity. Pacific Economic Review 12: 419-44. [CrossRef]
Johansen, Sø ren, and Katarina Juselius. 1992. Testing structural hypotheses in a multivariate cointegration analysis of the ppp and the uip for uk. Journal of Econometrics 53: 211-44. [CrossRef]
Johansen, Sø ren, Katarina Juselius Roman Frydman, and Michael D. Goldberg. 2010. Testing hypotheses in an $\mathrm{i}(2)$ model with piece-wise linear trends. an analysis of the persistent long swings in the $\mathrm{dmk} / \$$ rate. Journal of Econometrics 158: 117-29. [CrossRef]
Jorion, Phillippe. 2009. Risk management lessons from the credit crisis. European Financial Management 15: 923-33. [CrossRef]
Juselius, Katarina, and Katrin Assenmacher. 2017. Real exchange rate persistence and the excess return puzzle: The case of switzerland versus the us. Journal of Applied Econometrics 32: 1145-55. [CrossRef]
Juselius, Katarina, and Ronald MacDonald. 2004. International parity relationships between the usa and japan. Japan and the World Economy 16: 17-34. [CrossRef]
Juselius, Katarina, and Joshua Stillwagon. 2018. Are outcomes driving expectations or the other way around? an i(2) cvar analysis of interest rate expectations in the dollar/pound market. Journal of International Money and Finance 83: 93-105. [CrossRef]
Juselius, Katarina. 1992. Domestic and foreign effects on prices in an open economy: The case of denmark. Journal of Policy Modeling 14: 401-28. [CrossRef]
Juselius, Katarina. 1995. Do the purchasing power parity and the uncovered interest rate parity hold in the long run?-an application of likelihood inference in a multivariate time series model. Journal of Econometrics 69: 211-40. [CrossRef]
Juselius, Katarina. 2014. Testing for near $\mathrm{i}(2)$ trends when the signal-to-noise ratio is small. Economics-The Open-Access, Open-Assessment E-Journal 8: 1-30. [CrossRef]
Juselius, Katarina. 2017a. A Cvar Scenario for a Standard Monetary Model Using Theory-Consistent Expectations. Discussion Papers 17-08. Copenhagen: Department of Economics, University of Copenhagen.
Juselius, Katarina. 2017b. Using a theory-consistent cvar scenario to test an exchange rate model based on imperfect knowledge. Econometrics, MDPI, Open Access Journal 5: 1-20.
Knight, Frank Hyneman. 1921. Risk, Uncertainty and Profit. Boston: Houghton Mifflin, vol. 31
Lee, Byung-Joo. 2013. Uncovered interest parity puzzle: Asymmetric responses. International Review of Economics and Finance 27: 238-49. [CrossRef]
Levich, Richard M. 1985. Empirical studies of exchange rates: Price behavior, rate determination and market efficiency. In Handbook of International Economics. Edited by Ronald W. Jones and Peter B. Kenen. Amsterdam: Elsevier, vol. 2, pp. 979-1040.

Lewis, Karen K. 1995. Puzzles in international financial markets. In Handbook of International Economics. Edited by Gene M. Grossman and Kenneth Rogoff. Amsterdam: Elsevier, vol. 3, pp. 1913-17.
Liu, Wei, and Alex Maynard. 2005. Testing forward rate unbiasedness allowing for persistent regressors. Journal of Empirical Finance 12: 613-28. [CrossRef]
Lothian, James R., and Liuren Wu. 2011. Uncovered Interest- Rate Parity Over the Past Two Centuries. Journal of International Money and Finance 30: 448-73. [CrossRef]
MacDonald, Ronald, and Jun Nagayasu. 2015. Currency Forecast Errors and Carry Trades at Times of Low Interest Rates: Evidence From Survey Data on the Yen/Dollar Exchange Rate. Journal of International Money and Finance 53: 1-19. [CrossRef]
Mark, Nelson C., and Yangru Wu. 1998. Rethinking deviations from uncovered interest parity: The role of covariance risk and noise. Economic Journal 108: 1686-786. [CrossRef]
Maynard, Alex. 2003. Testing for forward-rate unbiasedness: On regression in levels and in returns. Review of Economics and Statistics 85: 313-27. [CrossRef]
McDowell, Allen. 2004. From the help desk: Seemingly unrelated regression with unbalanced equations. The Stata Journal 4: 442-48. [CrossRef]
Melvin, Michael, and Mark P. Taylor. 2009. The crisis in the foreign exchange market. Journal of International Money and Finance 28: 1317-30. [CrossRef]
Menkhoff, Lukas, Lucio Sarno Maik Schmeling, and Andreas Schrimpf. 2012. Carry trades and global foreign exchange volatility. Journal of Finance 67: 681-718. [CrossRef]
Moore, Michael J., and Maurice J. Roche. 2012. When does uncovered interest parity hold? Journal of International Money and Finance 31: 865-79. [CrossRef]
Nelson, Charles R., and Myung J. Kim. 1993. Predictable stock returns: The role of small sample bias. Journal of Finance 48: 641-61. [CrossRef]
Obstfeld, Maurice, Rogoff, Kenneth S., and Wren-Lewis Simon. 1996. Foundations of International Macroeconomics. Cambridge: MIT Press, vol. 30.
Olmo, Jose, and Keith Pilbeam. 2011. Uncovered interest parity and the efficiency of the foreign exchange market: A re-examination. International Journal of Economics and Finance 16: 189-204. [CrossRef]
Perron, Pierre. 2006. Dealing with structural breaks. In Palgrave Handbook of Econometrics, Vol. 1: Econometric Theory. Edited by Kerry Patterson and Terence C. Mills. London: Palgrave Macmillan, pp. 278-352.
Pettenuzzo, Davide, and Alan Timmermann. 2011. Predictability of stock returns and asset allocation under structural breaks. Journal of Econometrics 164: 60-78. [CrossRef]
Phillip, Bacchetta, and Eric van Wincoop. 2010. Infrequent portfolio decisions: A solution to the forward discount puzzle. American Economic Review 3: 870-904.
Sarno, Lucio. 2005. Towards a solution to the puzzles in exchange rate economics: Where do we stand? Canadian Journal of Economics 38: 673-708. [CrossRef]
Stambaugh, Robert F. 2006. Predictive regressions. Journal of Financial Economics 54: 375-421. [CrossRef]
Zhu, Zhen. 2002. Time-varying forward bias and the expected excess return. Journal of International Financial Markets, Institutions and Money 12: 119-37. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## MDPI

St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 616837734
Fax +41 613028918
www.mdpi.com

## Econometrics Editorial Office

E-mail: econometrics@mdpi.com
www.mdpi.com/journal/econometrics
MDPI

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel: +41616837734
www.mdpi.com


[^0]:    1 Sadly, Clive left us already in 2009, all too early.

[^1]:    2 Google scholar records 18132 citation in November 2020.

[^2]:    ${ }^{3}$ Such assumptions are, for example, long-run proportionality, exogeneity and endogeneity status, expectation formation, and so on.

[^3]:    4 Kongsted (2005) shows that under long-run price homogeneity it is possible to transform nominal money, $m$; income, $y$; and prices, $p$ to real money, $m-p$, real income, $y-p$, and inflation, $\Delta p$ without loss of information.

[^4]:    5 Many years later I revised my thinking on this: inflation is in fact a monetary problem, but after deregulation of capital movements, it is asset price inflation and house price inflation and not CPI inflation that react strongly on excess liquidity.

[^5]:    6 More than twenty years later, Johansen and Juselius (2014) showed that the invariance property was also valid for common stochastic trends, albeit in a slightly more complicated way. While the labeling of the estimated shocks changes when the information set is expanded, the stochastic trend(s) of the small model remain the same in the expanded model. Assume, for example, that the Danish bond rate is strongly exogenous in a domestic VAR model for Denmark and, thus, a common stochastic trend. Now, if the shocks to the Danish bond rate originate from shocks to the German bond rate and the German bond rate is added to the model, then the German rate becomes the "new" common stochastic trend. However, it is basically the same as the old one.
    7 Johansen et al. (2010) subsequently report a full-fledged $\mathrm{I}(2)$ analysis.

[^6]:    8 The variables in the figure are defined as follows: $w$ is nominal manufacturing wages, $p_{c}$ is consumer prices, $p_{y}$ is the price of output, $c$ is productivity, $u$ is unemployment, $R_{b}$ is the long-term bond rate, $R_{d}$ is the deposit rate, $\Delta p$ is CPI inflation, $m$ is money stock, $y$ is real GDP, $q$ is the real exchange rate, a superscript ${ }^{f}$ stands for foreign country, and a superscript * for an equilibrium value. A more detailed account can be found in Juselius (2006, Part VI).

[^7]:    9 In the dollar euro market, around $75 \%$ of all transactions are purely speculative. The remaining $25 \%$ are related to the trade in exports and imports but, due to forward contracts, future expectations play a significant role.

[^8]:    10 Motivated by the collapse of the Japanese real estate bubble a few years after the Finnish crisis, Richard Koo (2010) published his first book on balance sheet recessions.

[^9]:    11 The division into groups depended on whether foreign aid and the macro-economy-measured by GDP, investment, private consumption, and government expenditure-(1) had been unrelated in the long run; (2) whether aid had no long-run effect on the macro-economy-tested as a unit vector in $\alpha$-but the latter had been influencing aid; (3) whether aid has been exogenous with respect to the macro economy and finally; or (4) whether aid and the macro-economy have been tied together in an interdependent relationship.

[^10]:    12 This, in my view, is a sure proof that the person in question has never performed a proper CVAR analysis. Hundreds of students in the Copenhagen summer schools, who have struggled to make a well-specified CVAR deliver results in accordance with their favorite economic model-often without success-would certainly nod in agreement.

[^11]:    1 Usually orthonormality is assumed. This is convenient but not necessary in the present paper.

[^12]:    2 To our knowledge, the present paper is the first to study cointegration and error correction representations for $I(1)$ singular vectors, the factors of $I(1)$ dynamic factor models in particular. An error correction model in the DFM framework is studied in (Banerjee et al. 2014, 2017). However, their focus is on the relationship between the observable variables and the factors. Their error correction term is a linear combination of the variables $x_{i t}$ and the factors $\mathbf{F}_{t}$, which is stationary if the idiosyncratic components are stationary (so that the $x^{\prime}$ s and the factors are cointegrated). Because of this and other differences their results are not directly comparable to those in the present paper.

[^13]:    3 In the square case, $r=q$, Assumption 3 holds if and only if $\mathbf{M}(z)$ is unimodular.
    4 If $z^{*}$ is a zero of $\mathbf{M}(z)$, multiply $\mathbf{M}(z)$ by an invertible $r \times r$ matrix $\mathbf{Q}_{z^{*}}$ such that $z^{*}$ is a zero of, say, the first row of $\mathbf{Q}_{z^{*}} \mathbf{M}(z)$. Then multiply by the $r \times r$ diagonal matrix with $\left(z-z^{*}\right)^{-1}$ in position $(1,1)$ and unity elsewhere on the main diagonal. Iterating, all the zeros of $\mathbf{M}(z)$ are removed.

[^14]:    5 Multiplying both sides of (A3) by $(1-L)$ and using (A2), we obtain $\tilde{\mathbf{A}}(L) \mathbf{S}(L)^{-1} \mathbf{B}(L) \mathbf{u}_{t}=(1-L) \tilde{\mathbf{B}}_{t}$. Comparing the spectral densities of right- and left-hand terms, it is easy to prove that $\tilde{\mathbf{u}}_{t}$ must be a $q$-dimensional, nonsingular white noise and the rank of $\tilde{\mathbf{B}}$ must be $q$.

[^15]:    1 Please note that the original contribution to the estimation of cointegrating relationship has been least squares estimation in a non- or semi-parametric regression setting, see, e.g., Engle and Granger (1987). A recent survey of regression-based cointegration analysis is provided by Wagner (2018).

[^16]:    2 The complexity of these inter-relations is probably well illustrated by the fact that only Jensen (2013) notes that "even though the $\mathrm{I}(2)$ models are formulated as submodels of $\mathrm{I}(1)$ models, some $\mathrm{I}(1)$ models are in fact submodels of $\mathrm{I}(2)$ models".
    3 The literature often uses VAR models as approximations, based on the fact that VARMA processes often can be approximated by VAR models with the order tending to infinity with the sample size at certain rates. This line of work goes back to Lewis and Reinsel (1985) for stationary processes and was extended to (co)integrated processes by Saikkonen (1992), Saikkonen and Luukkonen (1997) and Bauer and Wagner (2005). In addition to the issue of the existence and properties of a sequence of VAR approximations, the question whether a VAR approximation is parsimonious remains.

[^17]:    4 Below we often use the term "likelihood" as short form of "likelihood function".
    5 We are confident that this dual usage of notation does not lead to confusion.

[^18]:    6 Our definition of VAR processes differs to a certain extent from some widely used definitions in the literature. Given our focus on unit root and cointegration analysis we, unlike Hannan and Deistler (1988), allow for determinantal roots at the unit circle that, as is well known, lead to integrated processes. We also include deterministic components in our definition, i.e., we allow for a special case of exogenous variables, compare also Remark 2 below. There is, however, also a large part of the literature that refers to this setting simply as (cointegrated) vector autoregressive models, see, e.g., Johansen (1995) and Juselius (2006).
    7 Of course, the statistical properties of the parameter estimators depend in many ways on the deterministic components.

[^19]:    8 The set $V_{p}$ is endowed with the pointwise topology $T_{p t}$, defined in Section 3. For now, in the context of VAR models, it suffices to know that convergence in pointwise topology is equivalent to convergence of the VAR coefficient matrices $a_{1}, \ldots, a_{p}$ in the Frobenius norm.
    9 Please note that in case of restricted estimation, i.e., zero restrictions or cross-equation restrictions, OLS is not asymptotically equivalent to PML in general.

[^20]:    10 A similar property holds for $V_{p, r}^{R R R}$ being a "thin" subset of $V_{p}^{O L S}$. This implies that the probability that the OLS estimator calculated over $V_{p}^{O L S}$ corresponds to an element $V_{p, r}^{R R R} \subset V_{p}^{O L S}$ is equal to zero in general.

[^21]:    11 Below Example 3 we clarify how these indices are related to the state space unit root structure defined in Bauer and Wagner (2012, Definition 2) and link these to the dimensions of the cointegrating spaces in Section 5.2.
    12 Uniqueness of realizations in the VAR case stems from the normalization $m(z) b(z)=I_{s}$, which reduces the class of observationally equivalent VAR realizations of the same transfer function $k(z)=a(z)^{-1} b(z)$, with $b(z)=I_{s}$, to a singleton.
    13 The pair $(a(z), b(z))$ is left coprime if all its left divisors are unimodular matrices. Unimodular matrices are polynomial matrices with constant non-zero determinant. Thus, pre-multiplication of, e.g., $a(z)$ with a unimodular matrix $u(z)$ does not affect the determinantal roots that shape the dynamic behavior of the solutions of VAR models.

[^22]:    14 When using the echelon canonical form, the partitioning is according to the so-called Kronecker indices related to a basis selection for the row-space of the Hankel matrix corresponding to the transfer function $k(z)$, see, e.g., Hannan and Deistler (1988, chp. 2.4) for a precise definition.

[^23]:    15 Here and below we will only consider state space systems in so-called innovation representation, with the same error in both the output equation and the state equation. Since every state space system has an innovation representation this is no restriction, compare Aoki (1990, chp. 7.1).

[^24]:    16 The definition of cointegrating spaces as linear subspaces allows to characterize them by a basis and implies a well-defined dimension. These advantages, however, have the implication that the zero vector is an element of all cointegrating spaces, despite not being a cointegrating vector in our definition, where the zero vector is excluded. This issue is well-known of course in the cointegration literature.

[^25]:    1 Here somewhat sloppily we use the same symbols for processes and their realizations.

[^26]:    2 Note that $\alpha=\left[I_{p_{1}}, 0\right]^{\prime}$, and thus $\Omega_{1 . c}=\left(\left[\Omega^{-1}\right]_{11}\right)^{-1}=\left(\alpha^{\prime} \Omega^{-1} \alpha\right)^{-1}=\left(\alpha^{\prime} g(1)^{\prime} \Sigma_{\epsilon}^{-1} g(1) \alpha\right)^{-1}=\left(\Phi_{1}^{\prime} \Sigma_{\epsilon}^{-1} \Phi_{1}\right)^{-1}$.

[^27]:    3 In this appendix processes whose dimension depends on the choice of $h$ are denoted using upper case letters neglecting the dependence on $h$ in the notation otherwise for simplicity.

[^28]:    4 Contrary to the usual Johansen notation we use $\Sigma_{\varepsilon}$ as the noise covariance and $\Omega$ as the variance of the Brownian motion corresponding to $\left(u_{t}\right)_{t \in \mathbb{Z}}$. Thus some of the formulas in this part show 'unusual' form.

[^29]:    5 A nice selection is such that if $h(i, j)$ is contained in the selection, then also $h(l, j)$ are contained for all $0<l<i$.

[^30]:    1 We use the Choleski factorization to ensure positive definiteness of the covariance matrices $\Omega_{u}, \Lambda$ and $\Omega_{\Phi}$. Thus, we estimate the parameters $B, A, \Omega_{u}=C_{u} C_{u}^{\prime}, \mu, \Phi, \Omega_{\Phi}=C_{\Phi} C_{\Phi}^{\prime}$ and $\Lambda=C_{\Lambda} C_{\Lambda}^{\prime}$ using Algorithm 2 and transform the covariances to the original parametrization. We obtain standard errors via the $\delta$-method.

[^31]:    2 Because we initialize at the true parameter value, the parameter sequences stabilize within the first 100 realizations. Using $K=600$ iterations is sufficient to reduce the impact of the approximation error.
    3 The choice of weight matrix is based on hand-tuning the convergence speed of Algorithm 2 by running a small number of trial-and-error runs with $N=50$ particles and constant step size $\gamma=1$.

[^32]:    4 Obtained via a Bloomberg LP Terminal using the ticker codes 'GDBR10 Index' and 'GGGB10YR Index'.

[^33]:    5 The difference between computing the classic standard errors with $N=1000$ and $N=10,000$ particles is negligible.
    6 Particle filter-based approximate robust standard errors have been suggested in Doucet and Shephard (2012), but we do not pursue this idea further in the present context.

[^34]:    1 Following Nielsen and Nielsen (2014), the choice of restrictions is of no importance for the resulting forecasts. Other normalizations could be considered; however, this gives an intuitive interpretation of $\alpha_{x}$.
    2 We restrict the data to 1950 and onwards as this removes outliers, and we avoid structural changes in the exposure; see Lee and Miller (2001) and Booth et al. (2002). To avoid uncertainty about the death rates, due to a few observations, we further restrict the ages to cover the ages 0 to 95 as is standard in the mortality forecasting literature.

[^35]:    3 This is also called the Strehler-Mildvan correlation due to Strehler and Mildvan (1960).

[^36]:    4 This is similar to Lee and Carter (1992) who assumed a homoskedastic error term for the LC model. The i.i.d. homoscedasticity assumption is necessary for the analysis of the present paper, but the assumption may be critical in certain cases-see, e.g., Doz et al. (2011).

[^37]:    5 This corresponds to the partial correlation squared between the fitted and observed values.

[^38]:    6 The MCS approach is implemented via the Ox-package Mulcom 3.0 by Hansen and Lunde (2014) in Oxmetrics 7, see Doornik (2013).
    7 For the case with only two models the forecast performance could be tested via the Diebold and Mariano (1995) test, which only allows for pairwise comparisons, whereas the MCS procedure allows for joint multiple model evaluation.

[^39]:    8 Note that we here use the period life expectancy (within year $t$ ), whereas the formula in Brouhns et al. (2002) computes the cohort life expectancy.
    9 These specifications have often been used in studies applying graduation laws of mortality-see Booth and Tickle (2008); McNown and Rogers $(1989,1992)$.
    10 In preliminary experiments, we also found this specification to give a better forecast performance compared with using other ARIMA models.
    11 The factors are estimated using weighted principal components in the R package Demography-see Hyndman and Ullah (2007) and Hyndman et al. (2014) for further details. All other models are estimated using own codes and the packages 'tsDyn', 'VARS', and 'Forecast' in R (R Core Team 2015) by Pfaff (2008); Stigler (2010) and Hyndman (2015).

[^40]:    1 For discussions of various approaches to causality in macroeconomics and macroeconometrics, see (Hoover 2001, 2008, 2012).
    2 See (Hoover 2001, chp. 3). In appealing to an experimental metaphor, Simon followed in the footsteps of Haavelmo (1944), a foundational figure for Cowles Commission econometrics (see Hoover and Juselius 2015; Hoover 2014).

[^41]:    3 "Graphical" (or "graph-theoretic") causal modeling should be the preferred term, as the search methods do not require a Bayesian approach to statistics. For compact treatments of the approach and the basic algorithms, see Cooper (1999) and Demiralp and Hoover (2003).
    4 On the general methodology of modeling in relation to the CVAR see Hoover et al. (2008) and Hoover and Juselius (2015).

[^42]:    5 Hoover (1990; 2001, chps. 2 and 3) provides a detailed account of Simon's approach and of it generalization to nonlinear systems, including ones with cross-equation restrictions among the parameters.

[^43]:    6 See Cooper (1999), Spirtes et al. (2000, chps. 5 and 6), and Pearl (2009, chp. 2). The Tetrad software package implements Spirtes et al. (2000) algorithms, as well as additional algorithms, and can be downloaded from Carnegie Mellon University's Tetrad Project website: http://www.phil.cmu.edu/tetrad/.

[^44]:    7 This question is addressed from a philosophical point of view in Hoover (2015).

[^45]:    8 In general, calculation of the cointegrating vector is the equivalent of solving out the $T \mathrm{~s}$ from the long-run representation of Equation (2) in which we set $\Delta \mathbf{X}_{\mathbf{t}}$ and the error terms to zero; specifically the cointegrating vector is given as $\boldsymbol{\Phi}_{\mathbf{x T}}^{\prime} \boldsymbol{\Phi}_{\mathbf{X X}}$ The orthogonal complement, indicated by the subscript is defined for a full-rank $p \times r$ matrix $\mathbf{A}$, as a $p \times(p-r)$ matrix $\mathbf{A}_{\perp}$, such that $\mathbf{A}_{\perp}^{\prime} \mathbf{A}=\mathbf{0}$; see (Johansen 1995, p. 39).
    9 Row 3 of $\Pi$ in DGP 2 is simply the first cointegrating relation from the reduced form of DGP 1 when $T$ is latent, while Row 2 is the second. Row 1 is $(-0.01) \times$ the first cointegrating relation $+(-0.1) \times$ the second.
    10 The eigenvalues of $\mathbf{I}+\Pi$ are $0.70678 \pm 0.16146 i$, and 0.98643 .

[^46]:    11 An analogous case arises in the graph-theoretic search literature in the guise of fragile failures of faithfulness-i.e., failures of the estimated probability distributions to reflect all of the independence relationships implied by the graph of the DGP (Spirtes et al. 2000, p. 41; Pearl 2009, pp. 62-63; Hoover 2001, pp. 45-49, 151-53, 168-69).

[^47]:    12 The robustness of trend behavior in CVARs driven by exogenous, latent trends would explain why the trends estimated in CVARs are often robust to widening the data set and recommends Juselius's specific-to-general approach-once the trends can be characterized, then any new variable is either redundant or carries information with respect to a new trend (Juselius 2006, chp. 22; Johansen and Juselius 2014).

[^48]:    $13 \Psi_{X X}$ is assumed to be full rank because, were it reduced rank, then it would itself generate trends in the manner of DGP 2 in Section 3-a case that we have argued is possible, but unlikely, in actual economies and which, therefore, we rule out by assumption in this analysis.

[^49]:    14 The connection of weak exogeneity to the efficient estimation of $\beta$ might suggest that our notion approach is similar to LeRoy (1995) approach to causality (cf. Hoover 2001, pp. 170-74). An importance difference, however, is that while LeRoy defines causal orderings in terms of efficient estimation, we seek only the implications for a possible of the lack of error correction of a condition that incidentally guarantees efficient estimation.

[^50]:    15 The orthogonal complement for any matrix is not, in general, unique; but each admissible complement spans the same space and places zero rows in the same positions.

[^51]:    16 This is, as in similar cases, a generic claim and does not rule out that zero rows in $\alpha$ might occur for carefully chosen coefficients.

[^52]:    17 The rule refers to the DGP, so that an unobserved intermediate cause would appear to warrant the inference of a direct causal connection when only an indirect connection existed in the DGP. This implies that widening the data set might, in effect, open the "black box" and provide more refined information about causal mechanisms.

[^53]:    1 Bank of England Inflation Report, November 2009, p.47. See http:/ /www.bankofengland.co.uk/publications/Documents/ inflationreport/ir09nov.pdf.

[^54]:    2 Other than the proposal in Hendry (2001), archival search revealed two earlier, somewhat unrelated, uses of 'forediction': (1) by Richard Feynman in a letter to Enrico Fermi in 1951 (see CBPF-CS-012/97); and (2) in Stenner (1964).
    3 Bank of England Minutes of the Monetary Policy Committee, November 2009, p.8. See
    http:/ /www.bankofengland.co.uk/publications/minutes/Documents/mpc/pdf/2009/mpc0911.pdf.

[^55]:    4 See https:/ /bankunderground.co.uk/2015/11/20/how-did-the-banks-forecasts-perform-before-during-and-after-thecrisis/.

[^56]:    5 All of these indicator saturation methods are implemented in the Autometrics algorithm in PcGive: see Doornik (2009) and Doornik and Hendry (2013), which can handle more variables than observations using block path searches with both expanding and contracting phases as in Hendry and Krolzig (2005), and Doornik (2007). An R version is available at https: / / cran.r-project.org/web/packages/gets/index.html: see Pretis et al. (2016).

[^57]:    6 As noted above, the lagged impact of the policy change causes $x_{191}$ to overshoot, so $\tilde{x}_{i, t}$ is somewhat above $x_{t}$ over the forecast horizon, albeit a dramatic improvement over Figure 3 (I): using 2-periods to estimate the IC solves that.

[^58]:    Note: $L-G a u s s i a n ~ l i k e l i h o o d, ~ A I C ~=~-~ 2 ~ l o g ~ L ~+~ 2 × ~ n u m b e r ~ o f ~ f r e e ~ p a r a m e t e r s, ~ S C ~=~-~ 2 ~ l o g ~ L+\log T \times$ number of free parameters, $\mathrm{HQ}=-2 \log L+2 \times$ number of free parameters $\times \log (\log T)$. MSIAH-model with time-varying intercept, VAR slope coefficients and error covariance matrix; MSIH-model with time-varying intercept and error covariance matrix; MSH—model with time-varying error covariance matrix but time-invariant intercept and VAR slope coefficients.

[^59]:    1 We also tried a Markov process with $M=3$ volatility states but failed to get reasonable Gaussian ML estimates.

[^60]:    1 The data set is available from the authors on request.
    2 The co-editors' observation.

[^61]:    3 The code is written in Ox, Doornik (2018).

[^62]:    4 One internal inconsistency that arose was that the test of omission of the trend from the cointegrated system was significant. Dropping the trend, does however not damage the diagnostics, and it gives interpretable results. Keeping it in the cointegration space does not.

[^63]:    5 Additional results, based on a sample split in 2000(4), before both the change in monetary policy and the enlargement of the EU labour market, are available from the authors.

[^64]:    1 For review articles, see Froot and Thaler (1990); Lewis (1995); Chinn (2006); Engel (1996, 2014); Sarno (2005).
    2 See Burnside et al. (2011a); Gourinchas and Tornell (2004); Mark and Wu (1998); Phillip and van Wincoop (2010).

[^65]:    3 An IKE risk premium model is developed in (Frydman and Goldberg 2007, 2013a).
    4 Juselius (1995) was the first to present evidence of this positive equilibrium relationship. See also (Juselius 1992, 2014, 2017a, 2017b); Cavusoglu et al. (2020); Frydman and Goldberg (2007); Hoover et al. (2008); Johansen and Juselius (1992); Johansen et al. (2010); Juselius and MacDonald (2004). See Brunnermeier et al. (2008), and Menkhoff et al. (2012) for additional evidence of downside risk in currency markets. In stock markets, see Ang et al. (2012).
    5 For example, see Bansal and Dahlquist (2000); Chinn (2006); Flood and Rose (2002); Frankel and Poonawala (2010); Ito and Chinn (2007); Lee (2013).
    6 To account for greater irrationality, Burnside et al. (2009) develop a model in which informed speculators' access to private information matters more in developed countries. Burnside et al. (2011a) assumes that market participants systematically overreact to information about future inflation. Phillip and van Wincoop (2010) develop a model of rational inattention.
    7 The BF regression also suffers from bias due to the much greater persistence in the forward premium compared with exchange rate changes. See Baillie and Bollerslev (2000); Liu and Maynard (2005); Maynard (2003); Nelson and Kim (1993); Olmo and Pilbeam (2011); Stambaugh (2006).

[^66]:    8 Goldberg et al. (Forthcoming) also find this kind of pronounced instability in $\beta$ for three developed countries. See also Bansal (1997); Clarida et al. (2009); Frydman and Goldberg (2007); Lothian and Wu (2011); Zhu (2002), and Baillie and Cho (2014).

    9 The sample sizes are limited by the availability of forward rate data, which is more difficult to obtain.
    10 Covered interest parity provides a very close proxy of the forward rate for economies without capital controls when eurocurrency interest rates are used. See Levich (1985).

[^67]:    11 The estimation procedure follows McDowell (2004) and accounts for the unbalanced data by creating a block diagonal matrix of all the countries.
    12 Six of the developed euro-area countries' samples extend farther back in time than in Frankel and Poonawala's dataset. We could include the individual spot and forward rate series in the unbalanced SUR model, ending the samples in December 1998. However, the individual European currencies were bound together in an ERM before 1999, which involved monetary policy cooperation among countries. These countries are thus best viewed as a single region for the developed country group. Denmark was tied to the ECU/euro in an informal ERM over the entire sample and so we treat it like a euro-area country.
    13 In general, the exchange rate series for any of the ERM/euro countries could be used as our euro-area series. We chose the Austrian spot and forward rate series since this country's sample period is the longest among the euro countries (extending back to 1976).

[^68]:    14 We are indebted to the editors for this argument and the SAR example that follows.
    15 SAR bid-asked prices were taken from a Bloomberg terminal. Bid-asked spreads are larger in the forward market.
    16 An earlier version of the paper included the tightly pegged USD regimes in the SUR analysis. The results of this analysis (which are available on request) are slightly more favorable to our main arguments.

[^69]:    17 The WMR forward rate data that we use for a few of the developed countries and many of Frankel and Poonawala's developing countries start one or more observations after their sample begins in December 1996. The full sample dates for each country in our panel are reported in column 2 of Tables 1 and 2.

[^70]:    18 As in Frankel and Poonawala (2010), the balanced sample starts in October 1997.
    19 Frankel and Poonawala drop Indonesia from the SUR analysis because their sample ends in February 2002. In order to facilitate a direct comparison, we also drop Indonesia from the analysis.
    20 We estimate the pooled model as unbalanced, unlike Frankel and Poonawala. This enables us to keep all developing countries in the analysis. The unbalanced SUR's time period is the same as in Frankel and Poonawala (2010).

[^71]:    21 The Wald test has finite sample limitations. However, it is a sensible choice here given that we do not have to estimate multiple models unlike other commonly employed multiple restriction tests. This is particularly advantageous because the unbalanced SUR model is computationally expensive compared to a balanced SUR model. The computational demands increase considerably in the model with subperiods as discussed below.
    22 To carry out the test, we create a dummy variable for the Frankel and Poonawala sample period. We add this dummy, and interaction terms with the remaining regressors, and estimate another unbalanced SUR model. The Wald statistic provides a test of the joint significance of these terms. In order to distinguish the effects of the extended sample period and list of countries, we repeat the same procedure, but limit it to the Frankel and Poonawala (2010) countries only. We find a significant difference between the two samples. These results are available upon request.

[^72]:    23 For example, see Engel (1996); Frydman and Goldberg (2007); Lewis (1995); Mark and Wu (1998).
    24 See also Ahmad et al. (2012), who finds that the Asian financial crisis triggered structural change in a panel of Asian-Pacific countries.
    25 The testing procedure proceeds equation by equation and thus ignores cross-country correlations in the data.
    ${ }^{26}$ We set the trimming level of the tests to $5 \%$, as opposed to the commonly employed $15 \%$. This decision allows for a wider portion of the sample to be considered in the test and relaxes the limit on the maximum number of allowed breaks. Relaxing this limit is important because we would expect many breaks for countries with the longest time series.
    27 We also perform supF tests, which confirm our break number and dating results.

[^73]:    28 We also estimate an unbalanced SUR model without the one-year subperiod restriction for the full panel. A few slope estimates become very large. However, the estimates for the common/comparable subperiods are close in magnitude. These results are available upon request.
    29 Alternative solutions to this problem include increasing the trimming parameter and/or imposing restrictions on the estimated coefficients, neither of which is suitable for our economic application. The former would limit the number of breaks, whereas there is no obvious bound to impose with the latter.
    30 Johansen et al. (2010) and Juselius (2014) find that interest rate differentials (and thus the forward premium) are near I(2). Bai and Perron (1998) suggest modeling a dynamic context by either adding lagged values to the regression or employing a nonparametric correction. Deng and Perron (2008) show in the context of other structural change tests that a dynamic specification helps address the autorrelated-errors problem.

[^74]:    31 Bansal (1997); Clarida et al. (2009); Frydman and Goldberg (2007); Lothian and Wu (2011); Moore and Roche (2012); Zhu (2002), and Baillie and Cho (2014) also report negative and positive estimates of subperiod biases.

[^75]:    32 See Baillie and Cho (2014); Brunnermeier et al. (2008); Melvin and Taylor (2009), and Daniel et al. (2017) for additional evidence of this time dependency and riskiness.
    33 The Wald tests in Table 9 consider the inidividual subsample biases, negative and positive, for developed and developing countries. The test is conducted under the null that the negative and positive biases for the two country groups are equal at mean.

[^76]:    34 Bekaert and Hodrick (1993) report that the large negative biases found for developing countries stem largely from behavior in the 1980s.
    35 Each country's sample begins with the observation right after the first break date in the 1990s. For example, the first break date in the 1990s for Australia is September 1993 (see Table A1). The truncated sample for this country therefore begins in October 1993.
    36 Goyal and Welch (2008) report similar results for linear models of stock returns.

[^77]:    37 In currency markets, see Goldberg and Frydman (1996a, 1996b); Ahmad et al. (2012); Beckmann et al. (2006); Melvin and Taylor (2009), and Goldberg et al. (Forthcoming). In stock markets, see Pettenuzzo and Timmermann (2011); Frydman and Goldberg (2011); Ang and Timmermann (2012), and Frydman et al. (2015).
    38 Brunnermeier et al. (2008), Daniel et al. (2017) and others find that carry trade returns are highly negatively skewed, which gives a measure of what they call âcrash or downside ârisk. These studies examine only developed-country markets.

