High-dimensional time series analysis

High-Dimensional Time Series Analysis: Unit Roots, Cointegration and Forecasting

E.J.J. Wijler

This book was typeset by the author using LaTeX.

Published by Universitaire Pers Maastricht
ISBN: 978 94 6380 745 6
Printed in The Netherlands by ProefschriftMaken

High-Dimensional Time Series Analysis: Unit Roots, Cointegration and Forecasting

DISSERTATION

to obtain the degree of Doctor at
Maastricht University,
on the authority of the Rector Magnificus,
Prof. dr. Rianne M. Letschert,
in accordance with the decision of the Board of Deans,
to be defended in public
on Thursday, 14th of January 2021, at 16.00 hours

by

Etienne Josepha Johannes Wijler

Supervisors

Prof. dr. A.W. Hecq

Dr. S.J.M. Smeekes

Prof. dr. J.R.Y.J. Urbain (†)

Assessment Committee

Prof. dr. F.C. Palm (Chair)

Prof. dr. H.P. Boswijk

Dr. M.C. Medeiros

Dr. I. Wilms

To Jos, Marlies, Suzanne, Guido and Daria.

Dedicated to the memory of Jean-Pierre Urbain, whose passion in research inspired me to take on this journey.

Acknowledgements

"As we express our gratitude, we must never forget that the highest appreciation is not to utter words, but to live by them."

- John F. Kennedy (1917-1963)

The work presented in this thesis would not have been possible without the guidance and support of many colleagues, friends and family. I would like to thank all of you whom have put up with me during this journey. As trying to enumerate all of you would surely result in me forgetting some of you, please know that your support has been sincerely appreciated.

There are, however, several people whom I have to mention explicitly. First of all, I want to thank the late Jean-Pierre Urbain, who inspired me to take on this long and ambitious journey. Jean-Pierre's ability to transfer his passion for research was truly remarkable and I consider myself lucky to have enjoyed the privilege of being one of his student. Jean-Pierre, I am forever indebted to you for igniting my passion for research. This thesis is dedicated to you.

I would not have arrived at my current destination without the help of my supervisor Stephan Smeekes. Stephan, I could not have wished for a better mentor throughout my PhD. It has been the greatest pleasure to explore the realm of high-dimensional statistics together with you. Moreover, for 'unne sjeng', the ability to converse in dialect, or rather an unusual mix of dialect and English jargon, is a luxury that cannot be understated. Thank you for the many lessons you have taught me. I look forward with great anticipation to our future collaboration together.

I am also greatly indebted to my second supervisor Alain Hecq, whose comments and reviews have frequently led to valuable insights and substantial improvements in

our work. Furthermore, I wish to thank Nalan Bastürk, Peter Boswijk, Jan van den Brakel, Marcelo Medeiros, Franz Palm and Ines Wilms for their willingness to form the reading committee and their effort to evaluate my thesis.

Next, a word of gratitude towards my two paranymphs, Hanno and Sean. Hanno, I met you as my tutor from my very first econometrics course and our friendship that has developed over the years has enriched my life academically, personally and culturally. I blame you for the constant classical music blasting through my speakers at the office. Perhaps the most fitting way to thank you, is to finally admit that the French horn is indeed the most beautiful instrument. Sean, we have known each other since high school, during which in many ways we seemed to be opposites. After unexpectedly becoming colleagues at Maastricht University, however, we discovered that we have many things in common after all, including our taste for the finer things in life. You have contributed to both my best nights out and my worst hang overs. I wouldn't want to miss any of them.

Finally, I would like to thank the people closest to me. Mom and dad, your endless love and support has shaped me into the person that I am today. Thank you for always being there for me, I consider myself lucky for having such caring parents. Making you pride has been a strong motivation throughout this journey and I hope this thesis can compensate for my rather rebellious teenage years. Suzanne, as my older sister I often looked at you as a role model. Your determination in life has been a great inspiration and has certainly contributed to my own success. Guido, thank you for the many laughs and valuable career-related advices. May MVV finally come out on top this season (or the next). The final words of gratitude go out to my beloved wife. Daria, you have given my life purpose beyond academic research. I deeply value your continuous support and your ability to re-energize me after a tiring day of work. Thank you for completing me.

Etienne Wijler
Maastricht, February 2020

Contents

Chapter 1

Introduction

"He who sees things grow from the beginning will have the finest view of them."

- Aristotle (384 - 322 BC)

In recent years, the availability of large datasets has become increasingly common in a wide variety of fields. Indeed, the term 'Big Data' is ubiquitous in both industry and academics, and especially prominent in the fields of computer science and econometrics (Diebold, 2012). While the term's exact definition remains ambiguous, and as a result is occasionally smirked upon by those strongly attached to the exact sciences, the general consensus is that 'Big Data' refers to the challenges of, and opportunities provided by, the analysis of increasingly large datasets. However, the origin and complexity of large datasets varies strongly across disciplines. As an example in the field of physics, the Large Hadron Collider (LHC) is the world's largest and most powerful particle accelerator, in which numerous detectors track the paths and energies of particles to provide digital summaries on collision events. The LHC produces roughly 25 Gigabytes of data per second, thereby posing a major challenge in terms of data processing. In time series econometrics, the field closest to this thesis, the growth in datasets commonly stems from increased institutional monitoring of financial and economic activity, and the measurement of variables at higher frequencies or lower levels of aggregation. While storage limitations are less troublesome for typical datasets in time series econometrics, their statistical analysis remains challenging as a result of data intricacies and the inability to manipulate the process that generates the data. Furthermore, ambitious model requirements such as the pursuit of simultaneous strong predictive power, interpretability and valid statistical inference, add an additional layer of complexity to the analysis. Particularly troublesome from

a statistical perspective are datasets in which the number of variables, henceforth referred to as the *dimension* of the dataset, is relatively large in comparison to the number of observational units. To distinguish this type of 'Big Data', we refer to such datasets as *high-dimensional* and the statistical methods tailored to the analysis of such datasets are referred to as high-dimensional statistics.

The literature on high-dimensional statistics is growing rapidly, and penalized regression has arisen as a promising method to model large datasets (e.g. De Mol et al., 2008; Kim and Swanson, 2014; Li and Chen, 2014). Penalized regression is a least-squares fitting procedure that imposes shrinkage to control the model complexity in high dimensions by penalizing the magnitude of estimated parameters. Contrary to ordinary least-squares regression (OLS), the added penalization enables estimation in high dimensions, even when the number of variables exceeds the number of observational units. Moreover, penalized regression is often praised for its ability to trade off a small increase in bias with a large reduction in variance of the estimates, a property that is particularly useful for prediction. In addition, certain variants of penalized regression, such as the 'least absolute shrinkage and selection operator' (lasso) by Tibshirani (1996), perform variable selection by setting coefficients equal to zero. As parsimonious models are easier to interpret, this property is especially relevant for applications aimed at describing relationships between variables in the data.

While early applications of penalized regression demonstrate favourable performance in high-dimensional settings (e.g. Hastie et al., 2008, Chapter 1), they are quite distant from those encountered in the field of time series econometrics. In time series analysis, issues such as cross-sectional correlation, serial dependence and, especially, non-stationarity, are known to affect the properties of statistical estimators. For example, spurious regression, which occurs when regressing unrelated unit root non-stationary variables on each other, invalidates standard forms of inference. Equally important is the related concept of cointegration, developed by Engle and Granger (1987), which describes how unit root non-stationary time series that share common stochastic trends may be linearly combined into a stationary process. Based on the plethora of tests for unit roots and cointegration proposed in the time series literature, along with the fact that Engle and Granger were awarded the Nobel Prize in economics for their work, it is hard to overstate the academic and practical relevance of these topics. Because the estimation procedure of penalized regression depends on a least-squares component, there is no a priori reason to believe that these estimators are unaffected by the (co)integration properties of the data. Clearly, application of penalized regression to (non-stationary) time series settings demands a separate analysis of its theoretical properties and empirical performance.

This thesis theoretically and empirically analyses penalized regression methods in realistic time series settings and develops a novel estimator tailored to high-dimensional applications based on (co)integrated datasets. The methods are analysed from an asymptotic perspective, with an emphasis on properties related to estimation accuracy and variable selection. In several empirical applications, the predictive performance of penalized regression methods is analysed and compared to popular alternative modelling procedures. The objective of the thesis is to validate the use of penalized regression to (non-)stationary time series applications, as well as to extend the toolbox of the applied time series researcher.

Let us now go into more detail. First, we formally introduce penalized regression and discuss several important concepts such as sparsity, selection consistency and the oracle property. Afterwards, we briefly review the most prominent challenges of time series analysis in high dimensions. Next, we motivate penalized regression as a potential solution to these challenges and highlight the contribution of the thesis with links to the following chapters. Finally, we discuss some limitations to this thesis and propose several interesting avenues for future research.

Notation

Throughout the thesis, we follow the notation proposed by Abadir and Magnus (2002) as closely as possible. In particular, a scalar is denoted by a lowercase letter (x), a vector by a boldface lowercase letter ($\boldsymbol{x}$) and a matrix by a boldface uppercase letter ($\boldsymbol{X}$). By convention, a vector is interpreted as a column-vector. Additional relevant notation is introduced separately in the consecutive chapters.

1.1 Challenges in High-Dimensional Time Series Analysis

The extraction of information from a collection of time series is central to time series econometrics, and much effort is devoted to accommodate for datasets of larger dimensions. Insightful examples of time series in econometrics are those of a financial analyst that uses daily closing prices of stocks to empirically verify the CAPM model, or of an economist considering monthly inflation rates to explore the effects of changes in economic policy. Classical time series analysis concerns the specification and estimation of models that best capture the dynamic features of the data, with a particularly important consideration being whether the time series at hand are integrated or stationary. Indeed, one of the first decisions a researcher faces is whether to

correct for possible unit root non-stationarity by differencing the data, or by adopting a model that explicitly incorporates the integrating properties. This is non-trivial in low dimensions and several additional challenges arise in the high-dimensional setting.

First, the process of pre-testing for unit roots is substantially more complicated in high dimensions. At the early stages of the model building process, the decision on the correct dynamic specification is commonly based on a procedure that tests each time series for the presence of a unit root. Consequently, among the first issue to arise in high dimensions, is that naively pre-testing a large number of individual time series quickly accumulates the probability of making a false rejection. The literature on multiple hypothesis testing proposes several solutions, often designed to control the family-wise error rate or the false discovery rate (see Romano et al., 2008b, for a review). However, the decision of which metric to focus on, as well as the preferred strategy by which to optimize this metric, is often unclear and data-dependent, as illustrated in Chapter 5. Furthermore, the impact of misspecification of the order of integration depends on the robustness and purpose of the subsequent estimation procedure. Accordingly, the effect of potential errors in the pre-testing procedure is an important consideration in this thesis.

Second, model estimation in high dimensions adds computational challenges. The use of simple least-squares routines without imposing additional regularization to control for model complexity exhausts the degrees of freedom and, consequently, provides inaccurate estimates. Shrinkage estimators solve this issue by regularizing the solutions, but can be computationally demanding when no analytic expression exists and numerical optimization is required. Indeed, computational simplicity is an important motivation behind our shrinkage estimator developed in Chapter 3.

Finally, classic theory for time series models is often not well-suited to high-dimensional applications. Popular time series models, such as the vector autoregressive model (VAR) for stationary data or the vector error-correction model (VECM) for integrated data, are typically motivated in a fixed-dimensional asymptotic framework; asymptotic results are derived under the assumption that the number of variables N is kept fixed while the time series dimension T diverges. Such a setting, however, is in stark contrast with high-dimensional datasets in which N is relatively large to T, resulting in poor quality asymptotic approximations. Thus, an asymptotic framework that accounts for the effect of dimensionality is required, and constitutes the central topic of Chapter 4.

Evidently, the challenges brought forward by the increasing dimensionality of modern datasets necessitate alternative modelling strategies. An approach that has long

been dominant in the time series econometrics literature consists of factor models, which rely on the assumption that the data is driven by a small number of unobserved common components (see Bai and Ng, 2008b, for an elaborate survey). However, one may believe that only a subset of the observed data is required for accurately explaining the variation in the variables of interest. Since factor models are incompatible with this philosophy, we consider the use of penalized regression methods as a solution instead.

1.2 Penalized Regression

In this section, we formally introduce the method of penalized regression. For illustrative purposes, assume that we observe a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, where $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})'$. Additionally, suppose we wish to use this sample to explain the variation in a dependent variable y_i, whose true (unobserved) data-generating process (DGP) is described by

$$y_i = \sum_{j=1}^{s} \beta_j x_{i,j} + \epsilon_i = \boldsymbol{\beta}' \boldsymbol{x}_i + \epsilon_i, \tag{1.1}$$

where $s < p$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_s, \mathbf{0}')'$ and ϵ_i is a random error term with $\mathbb{E}(\epsilon_i) = 0$. Furthermore, we use $\boldsymbol{\beta}_{S_\beta} = (\beta_1, \ldots, \beta_s)'$ to denote the support of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_{S_\beta^c}$ denotes its complements, i.e. a $(p-s)$-dimensional vector of zeroes. A characterizing feature of DGP (1.1) is that only a subset of all variable is relevant to explaining the variation in y_i. We refer to such DGPs as *sparse*.[1]

Penalized regression imposes regularization on top of the standard least squares fitting procedure to control for model complexity, thereby enabling application to high-dimensional datasets. The regularization is introduced through the addition of a penalty to the standard least squares objective function:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{i,j} \beta_j \right)^2 + P_\lambda \left(\boldsymbol{\beta} \right), \tag{1.2}$$

where $P_\lambda \left(\hat{\boldsymbol{\beta}} \right)$ represents the penalty function that regularizes model complexity by shrinking the coefficients. A large variety of penalty terms are proposed in the literature, and the estimators enjoy different properties depending on the specific form of

[1] Clearly, sparsity of a DGP only makes sense in reference to the data on which the analysis is conditioned.

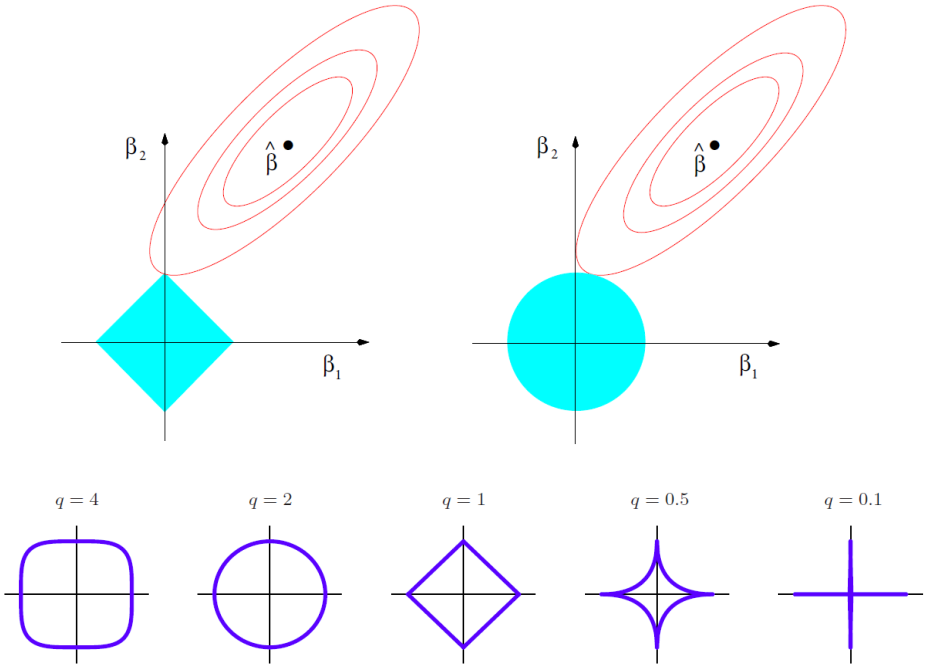

Figure 1.1: This figure corresponds to Figures 2.2 and 2.6 in Hastie et al. (2015). The first panel displays the estimation of the lasso (left) and ridge regression (right). The blue shaded area corresponds to the constraint regions $\|\boldsymbol{\beta}\|_1 \leq k$ and $\|\boldsymbol{\beta}\|_2^2 \leq k$, respectively. The red ellipses are the contours of the residual sum of squares. Panel 2 displays the contraint regions corresponding to $|\beta_1|^q + |\beta_2|^q \leq k$ for different values of q.

the penalty (e.g. Hastie et al., 2015). Most commonly, the penalty corresponds to a scaled L_q-norm, such as

$$P_\lambda(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_q, \tag{1.3}$$

where λ is a tuning parameter that regulates the degree of shrinkage and $\|\boldsymbol{\beta}\|_q = \left(\sum_{j=1}^p |\beta_j|^q\right)^{1/q}$. Among the most familiar variants are the lasso, which uses an L_1-norm, and ridge regression, which incorporates a squared L_2-norm. The use of an L_q-norm with $q \geq 1$, has the benefit that the objective function in (1.2) is convex, thereby simplifying computations and guaranteeing uniqueness of the minimizer. Alternatively, while being computationally more challenging, the use of an L_q-norm with $q \leq 1$ results in sparse estimates in which some coefficients are shrunken to be exactly

equal to zero.[2] An intuitive explanation for this sparsity inducing property is visualized in Figure 1.1. For an artificial dataset with $p = 2$, the first panel in Figure 1.1 displays the contours of the sum of squared residuals (red lines) and the constraint regions of different penalty functions (blue shaded areas). The solution that minimizes the objective function (1.2) is located at the point where the contours touch the boundary of the constraint region. It is intuitively clear that for sharp-cornered and diamond-shaped constraint regions, the solution is likely to lie at a corner point with one of the coefficients set equal to zero. As displayed in panel 2, such constraint regions correspond to L_q-norms with $q \leq 1$. The lasso, therefore, is a unique form of penalized regression in the sense that it relies on the only L_q-norm that induces sparsity while maintaining convexity of the objective function.

Fitting procedures that perform simultaneous estimation and variable selection are often desired to possess several attractive asymptotic properties. A requirement that is familiar from the fixed-dimensional literature is that of *estimation consistency*, i.e. the estimates converge in probability to the true values as the sample size grows:

$$\mathbb{P}\left(\left|\hat{\beta}_j - \beta_j\right| > \epsilon\right) \to 0, \tag{1.4}$$

for each $j = 1, \ldots, p$ as $n \to \infty$. It is important to note that (1.4) does not imply that for any finite sample some coefficients are in fact estimated as exactly zero; convergence in probability requires the estimated coefficients to grow closer to the true values with high probability, without necessarily ever being exactly equal to the true value. However, when the estimator is to be used as a variable selection device, a natural requirement is that the set of relevant variables is correctly identified with high probability when the sample size grows large. This is captured by the notion of *selection consistency*, which states that

$$\mathbb{P}\left(\left\{j : \hat{\beta}_j \neq 0\right\} = \{j : \beta_j \neq 0\}\right) \to 1, \tag{1.5}$$

as $n \to \infty$. Zhao and Yu (2006) introduce the stronger notion of *sign consistency*, which also requires the signs of the non-zero coefficients to be estimated correctly in the limit:

$$\mathbb{P}\left(\text{sign}\left(\hat{\boldsymbol{\beta}}\right) = \text{sign}\left(\boldsymbol{\beta}\right)\right) \to 1, \tag{1.6}$$

as $n \to \infty$, with (1.6) holding element-wise. Establishing estimation consistency and

[2]This variable selection property is especially relevant when the DGP is believed to be sparse, although we show in Chapter 2 that sparse methods may perform well in certain non-sparse settings.

selection consistency is an essential part of providing asymptotic justification for the use of penalized regression, provided that the assumptions under which these results hold are realistic for the specific application considered.

Remark 1.1. The probabilistic statements thus far presented rely on shorthand notation. Formally, we assume that $\{(\boldsymbol{x}_i', \epsilon_i)'\}_{i=1}^{\infty}$ is a sequence of $\left(\mathbb{R}^{p+1}, \mathscr{B}^{p+1}\right)$-valued random variables defined on some underlying probability space $(\boldsymbol{\Omega}, \mathscr{F}, \mathbb{P})$, such that $\hat{\boldsymbol{\beta}}(\omega) : \boldsymbol{\Omega} \to \mathbb{R}^p$ is a function of the events in this space. Then, for example, (1.6) ought to be interpreted as

$$\mathbb{P}\left(\bigcap_{j=1}^{p}\left\{\omega : \operatorname{sign}\left(\hat{\beta}_j(\omega)\right) = \operatorname{sign}(\beta_j)\right\}\right) \to 1,$$

as $n \to \infty$. As this notation may become unnecessarily technical at times, we rely on shorthand notation without reference to the underlying probability space when possible to do so without ambiguity.

Among all fitting procedures that deliver consistent estimation and selection, one may desire to choose the most accurate estimator. Since, by selection consistency, $\mathbb{V}\mathrm{ar}\left(\hat{\boldsymbol{\beta}}_{S_\beta^c}\right) = 0$ on a set with probability converging to one as $n \to \infty$, an efficiency argument is necessarily based on $\mathbb{V}\mathrm{ar}\left(\hat{\boldsymbol{\beta}}_{S_\beta}\right)$. Fan and Li (2001), and later Zou (2006), define the *oracle property* as a criteria by which to evaluate the optimality of a fitting procedure that performs simultaneous estimation and variable selection. Formally, the estimator $\hat{\boldsymbol{\beta}}$ possesses the oracle property if

1. $\mathbb{P}\left(\hat{\boldsymbol{\beta}}_{S_\beta^c} = \mathbf{0}\right) \to 1$, and
2. $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{S_\beta} - \boldsymbol{\beta}_{S_\beta}\right) \xrightarrow{d} \mathscr{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^*\right)$,

as $n \to \infty$, where $\boldsymbol{\Sigma}^*$ is the asymptotic variance of the ordinary least-squares (OLS) estimator applied directly to the true subset of relevant variables.[3] Intuitively, an estimator that possesses the oracle property consistently selects the correct subset of relevant variables and estimates their coefficients with the same efficiency as if the relevant variables were known beforehand. For some variants of penalized regression, such as the adaptive lasso introduced in Chapter 2, it is possible to derive this oracle property, although one typically needs to restrict the parameter space for such results to hold uniformly.

[3] In the context of penalized maximum-likelihood estimation, one can define $\boldsymbol{\Sigma}^*$ as the Cramer-Rao lower bound based on the true subset of relevant variables.

1.3 Penalized Regression in Time Series: Contribution of This Thesis

The properties of penalized regression described in Section 1.2 offer prospective solutions to the challenges in high-dimensional time series analysis laid out in Section 1.1. For example, on sparse DGPs, the estimation and selection consistency of the lasso allow for fast and efficient estimation in high dimensional settings without exhausting the degrees of freedom. Moreover, the asymptotic theory of penalized regression methods can be altered to accommodate for high-dimensional settings, thereby providing more realistic asymptotic approximations. These and other contributions of this thesis are summarized below.

In Chapter 2, we consolidate separately proposed lasso-type estimators[4] for stationary time series data and we systematically compare their predictive and selective performance in controlled settings, as well as on empirical applications. The analysis largely focusses on comparisons between a variety of penalized regression methods and factor models. The key insights are that penalized regression methods are more robust than factor models; they display superior predictive performance on sparse DGPs, while performing only marginally worse than factor models on DGPs with a factor structure. Moreover, when the idiosyncratic component of the factor model is not sufficiently 'well-behaved', penalized regression actually outperforms the factor models. In addition, we obtain some anecdotal evidence that lasso-type estimation in non-stationary setting may bring forecast improvements over traditional OLS estimators, but simultaneously observe a high sensitivity to the (co)integrating properties of the data in higher dimensions.

In recognition of the sensitivity of lasso-type methods in non-stationary settings, we develop an intuitive lasso-type estimator designed to properly take into account the (co)integrating properties of the data in Chapter 3. The estimator, referred to as the Single-equation Penalized Error Correction Selector (SPECS), relies on penalized estimation of a conditional error correction model, and is shown to posses the oracle property in a fixed-dimensional asymptotic framework. Importantly, this property holds without requiring any pre-testing for unit roots. Simulations demonstrate superior performance compared to alternatives that ignore the (co)integration properties and an empirical application in which we nowcast Dutch unemployment with the use of Google Trends confirms these findings.

[4] The term 'lasso-type estimators' loosely refers to variants of penalized regression that involve an L_1-penalty.

Chapter 4 extends the theory for SPECS to a high-dimensional asymptotic framework, allowing the cross-sectional dimensions of both short-run and long-run dynamics to diverge alongside the time series dimension. The results confirm that estimation and selection consistency are attainable in a high-dimensional setting, although the dimension and the convergence rate of the estimator are inversely related. Furthermore, the generality of the theoretical framework is restricted by the absence of knowledge on the behaviour of the minimum eigenvalues of high-dimensional sample covariance matrices containing integrated processes.

Following the advent of new high-dimensional methods that allow for direct application to non-stationary datasets, Chapter 5 reviews and compares two main high-dimensional modelling approaches: (i) identifying unit roots and transforming all data to stationarity versus (ii) explicitly modelling any unit roots and cointegrating relationship. We provide a detailed illustration of common pitfalls of unit root testing in high dimensions and evaluate methods designed to deal with issues such as poor size and power of unit root tests, as well as controlling appropriate error rates in multiple testing. In two empirical applications, we incorporate specialized factor models and penalized regression methods that accommodate both modelling approaches and we examine their comparative predictive performance. We find that no method of modelling cointegration arises as superior and that the potential gains from taking into account cointegration for forecasting remains data-dependent. We are led to conclude that model specification will always require careful consideration, although the practitioner benefits from access to an increasingly large set of reliable tools to model unit roots and cointegration.

Finally, we comment on some relevant topic this thesis does not consider. First, the methods in this thesis are solely motivated from the frequentist point of view. While many methods included in our comparative analyses have Bayesian counterparts (c.f. Park and Casella, 2008), the large collection of penalized regression methods and factor models prevents us from drawing from the large pool of Bayesion methods without losing focus on the main research questions. Second, the lasso can be seen as part of a larger class of estimators referred to as folded non-concave penalized maximum likelihood estimators (Fan and Li, 2001). While this class of estimators contains penalty functions that lead to attractive theoretical properties, such as the smoothly clipped absolute deviation (SCAD) penalty, many result in non-convex objective functions that lead to more complicated estimation procedures. Hence, we only take into account L_q-penalized regression with $q = 1, 2$. Third, we do not discuss post-model selection inference. It is now well-recognized that such inference is complicated by the issue of post-selection bias and numerous solutions, such as post-double selection

(Belloni et al., 2014) or the desparsified lasso (Van de Geer et al., 2014), are available. However, none of these approaches extend easily to general stationary time series settings, and extensions to the unit root setting are expected to be highly complicated. Nonetheless, the theoretical results of Chapter 3-4 may prove useful as intermediary results in the pursuit of uniformly valid post-model selection inference. We consider this an exciting avenue for future research.

Chapter 2

Macroeconomic Forecasting Using Penalized Regression Methods

"Recent advances in information technology make it possible to access in real time, at a reasonable cost, thousands of economic time series for major developed economies. This raises the prospect of a new frontier in macroeconomic forecasting, in which a very large number of time series are used to forecast a few key economic quantities, such as aggregate production or inflation."

- Stock and Watson (2002)

Abstract[†]

In this chapter, we investigate the suitability of lasso-type penalized regression techniques when applied to macroeconomic forecasting with high-dimensional data sets. We consider the performance of lasso-type methods when the true data generating process (DGP) is a factor model, contradicting the sparsity assumption underlying penalized regression methods. We also investigate how the methods handle unit roots and cointegration in the data. In an extensive simulation study we find that penalized regression methods are more robust to mis-specification than factor models, even if the underlying DGP possesses a factor structure. Furthermore, the penalized regression methods are demonstrated to deliver forecast improvements over traditional approaches when applied to non-stationary data containing cointegrated variables, despite a deterioration of the selective capabilities. Finally, we also consider an empirical application to a large macroeconomic U.S. dataset and confirm the competitive performance of penalized regression methods.

[†]This chapter is based on Smeekes and Wijler (2018b).

2.1 Introduction

In this chapter we provide a thorough analysis of the forecasting capabilities of penalized regression in macroeconomic conditions. We study the performance of these methods in a simulation study when the true DGP is a factor model and when the data contain stochastic trends and may be cointegrated. We also provide a systematic comparison with factor models, the mainstream method used in macroeconomic forecasting, using both Monte Carlo simulations and an empirical application to macroeconomic data.

Despite the vast size of the forecasting literature, comprehensive comparisons between factor models and penalized regression remain scarce. Traditionally, the majority of the forecasting literature seems to have implicitly assumed the prevalence of a latent factor structure in economic datasets and therefore has mainly considered the performance of methods based on factor estimation. While very popular in statistics, only recently L_1-penalized regression techniques, such as the lasso from Tibshirani (1996), are being explored as a viable alternative to traditional estimators such as low-dimensional VARs or factor models, in macroeconometrics. Applications in forecasting in particular show that the use of penalized regression, potentially in combination with traditional techniques such as principal components (PC), delivers promising performance (e.g Kim and Swanson, 2014; Garcia et al., 2017), though it is not yet really understood why. By providing a comprehensive study of penalized regression in 'adverse' macroeconomic conditions, we complement the existing literature with a fresh perspective on these methods and a direct link to factor models.

Specifically, we address the apparent contradiction between the premise of forecasting with shrinkage estimators to identify a small subset of variables responsible for the variation in the dependent variable and the assumption that the variation in the dependent variable is best explained through aggregates of all available time series. The good empirical performance of penalized regression methods despite this contradiction gives rise to a number of practically relevant questions; (1) Is the common factor assumption really valid in practice? (2) Are the results due to sample-dependent data idiosyncrasies? (3) Are other mechanisms at play such as an inherent robustness of shrinkage estimators to alternative DGP specifications?

We aim to shed light on these previously unexplored questions by conducting a detailed simulation study in which we compare the performance of a selection of the most popular and well understood variants of L_1-shrinkage estimators and factor extraction methods. The novelty in these simulations comes from the wide range of

DGPs considered, chosen such that no method is consistently favoured over another based on a priori expectations and to closely resemble the types of data that occur in empirical applications. The former goal is maintained through varying both the presence of common factors in the data as well as the degree of sparsity in the parameter space, while the latter goal is maintained through introducing levels of non-sphericity frequently encountered in empirical work.[1] In addition, we explore the potential of penalized regression in the non-stationary setting by generating a number of time series containing unit roots, some of which are cointegrated, and employ penalized regression directly on these series without any form of preprocessing. We complement the simulations with a comparison of the pseudo out-of-sample forecasting performance on a recently updated U.S. macroeconomic dataset available through the Fred-MD database (McCracken and Ng, 2016).

The results show that penalized regression performs remarkably well when there is at least some degree of sparsity in the parameter space and is relatively robust against alternative DGP specifications. Factor models perform slightly better than penalized regression when the predictors possess an approximate factor structure with low dependence in the errors, but their performance deteriorates substantially when increasing the level of non-sphericity in the idiosyncratic component. Penalized regression naturally does better than factor models on sparse DGPs, but more surprisingly also provides forecast improvements on DGPs containing a factor structure with strongly serially and cross-sectionally correlated idiosyncratic components. In addition, penalized regression shows promising results on cointegrated data, producing substantially lower forecast errors compared to standard OLS, despite failing to identify the exact cointegrating vector at relatively high frequencies. Finally, the empirical application highlights that the forecast performance differentials between factor-based methods and shrinkage methods are sensitive to the target variable being forecast.

Our contribution complements the vast existing macroeconomic forecasting literature that is dominated by methods that exploit a latent factor structure, such as static factor models (e.g. Stock and Watson, 2002a,b; Bai and Ng, 2008a), dynamic factor models (Eickmeier and Ziegler, 2008; Forni et al., 2005a, 2018; Doz et al., 2012), weighted principal components (Boivin and Ng, 2006), sparse principal components (Kristensen, 2017) or factor augmented vector autoregressions (Bernanke et al., 2005b; Pesaran et al., 2011; Bai et al., 2016). The conjecture that a small set of factors drives the variation in economic time series finds strong support through impressive forecasting performance of factor models on macroeconomic datasets from the U.S. (Stock

[1] Throughout this chapter the term non-sphericity refers to the presence of cross-sectional and/or serial correlation in the idiosyncratic component of a data generating process.

and Watson, 2002a, 2012), the U.K. (Artis et al., 2005) and the Euro area (Marcellino et al., 2003). Spurred by theoretical developments, such as the extension of the adaptive lasso to general time series frameworks by Medeiros and Mendes (2016), L_1-penalized regression has gained more appeal and the body of applied literature taking into account these shrinkage estimators has grown considerably. Recent work covers penalized regression (Gelper and Croux, 2008; De Mol et al., 2008; Kim and Swanson, 2014; Li and Chen, 2014), reduced-rank vector autoregressions (Bernardini and Cubadda, 2015), Bayesian vector autoregressions (Bańbura et al., 2010) and penalized vector autoregressions (Hsu et al., 2008; Callot and Kock, 2014; Kascha and Trenkler, 2015; Barigozzi and Brownlees, 2019). While some include a direct comparison between at least some form of factor models and penalized regression and demonstrate predictive capabilities of L_1-penalized regression that are competitive to traditional factor models, the analysis is typically based on empirical data or simulations that do not provide detailed insights into the sensitivity of each method to its underlying assumptions.

The remainder of this chapter is organized as follows. Section 2.2 describes the notation and reviews the methods considered. In section 2.3 we perform the simulation based analysis of the forecasting performance, followed by the empirical application in section 2.4. In section 2.5 we conclude and suggest a number of interesting avenues for future research.

2.2 Methods

Suppose a researcher is interested in predicting an economic time series h-steps ahead with information available through time $t = 1, \dots, T$. The researcher desires to include a pre-determined set of variables such as lags of the dependent variable or variables motivated through economic theory. In addition, she faces a large set of candidate variables that are potentially relevant to the dependent variable. This results in the following generic model:

$$y_{t+h} = \boldsymbol{w}_t'\boldsymbol{\beta}_w + \boldsymbol{x}_t'\boldsymbol{\beta}_x + \epsilon_{t+h} \tag{2.1}$$

where y_{t+h} is the scalar valued dependent variable to forecast and h is the forecast horizon. Furthermore, $\boldsymbol{w}_t$ is the $(p \times 1)$ predetermined vector of variables which the researcher requires to be in the model, $\boldsymbol{x}_t$ is the $(N \times 1)$ vector containing candidate variables that are potentially related to y_{t+h}, and ϵ_{t+h} is a disturbance term. The forecast of the response at time T is defined as $\hat{y}_{T+h|T} = \boldsymbol{w}_T'\hat{\boldsymbol{\beta}}_w + \boldsymbol{x}_T'\hat{\boldsymbol{\beta}}_x$. Letting $\boldsymbol{y} =$

$(y_{1+h}, \ldots, y_{T+h})'$, $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T)'$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)'$ and $\boldsymbol{\epsilon} = (\epsilon_{1+h}, \ldots, \epsilon_{T+h})$ the model can be rewritten as

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{\beta}_w + \boldsymbol{X}\boldsymbol{\beta}_x + \boldsymbol{\epsilon}. \tag{2.2}$$

When the number of variables in the candidate set $\boldsymbol{X}$ is large relative to the number of available observations, modelling the dependent variable as a linear combination of all candidate variables will amount to the estimation of a large number of parameters and is likely to result in a large forecasting variance. For example, assuming the explanatory variables follow a Gaussian distribution, Stock and Watson (2006) show that the OLS forecast is normally distributed with a variance proportional to the number of variables included in the model divided by the total number of available observations. In the more extreme case where the cross-sectional dimension exceeds the time series dimension inverting the matrix of second moments becomes infeasible and as a result the OLS estimator does not have a (unique) solution. Accordingly, methods that perform regularization are required in order to obtain accurate forecasts and reliable model estimates in the high-dimensional setting.

The methods we consider can broadly be categorized as shrinkage estimators and factor models. Shrinkage estimators aim to reduce the forecast variance by shrinking the parameter estimates in the traditional linear model, possibly up to a point where some parameters are exactly equal to zero and, thus, removing the corresponding variables from the candidate set. Factor models, on the other hand, do not remove variables from the candidate set, but rather aim to reduce the dimensionality of the data by summarizing the data in relatively few factors with the hope of capturing the bulk of the variation in the candidate set. In the following section we formally introduce these methods and describe the mechanisms by which they estimate our generic model (2.1).

2.2.1 Shrinkage estimators

The shrinkage estimators employed in this chapter estimate the parameters according to the following objective function:

$$\begin{aligned}(\hat{\boldsymbol{\beta}}_w, \hat{\boldsymbol{\beta}}_x) = &\underset{(\boldsymbol{\beta}_w, \boldsymbol{\beta}_x)}{\arg\min} \sum_{t=1}^{T} (y_{t+h} - \boldsymbol{w}_t'\boldsymbol{\beta}_w - \boldsymbol{x}_t'\boldsymbol{\beta}_x)^2 \\ &+ \lambda \left[\alpha \sum_{j=1}^{N} \frac{|\beta_{x,j}|}{\omega_j} + (1-\alpha) \sum_{j=1}^{N} \frac{|\beta_{x,j}|^2}{\omega_j} \right],\end{aligned} \tag{2.3}$$

with different settings of $(\lambda, \alpha, \omega_j)$ leading to various well-established methods. We consider:

1. Ridge regression (ridge: $\lambda > 0, \alpha = 0, \omega_j = 1 \ \forall j$)
2. Lasso (las: $\lambda > 0, \alpha = 1, \omega_j = 1$),
3. Adaptive Lasso (adalas: $\lambda > 0, \alpha = 1, \omega_j = \left|\hat{\beta}_{Init,j}\right|$),
4. Elastic Net (en: $\lambda > 0, 0 < \alpha < 1, \omega_j = 1 \ \forall j$), and
5. Adaptive Elastic Net (adaen: $\lambda > 0, 0 < \alpha < 1, \omega_j = \left|\hat{\beta}_{Init,j}\right|$),

where $\hat{\beta}_{Init,j}$ is an initial estimate such as the OLS or ridge coefficient. All methods impose shrinkage ($\lambda > 0$) that enables model estimation in situations where the number of potentially relevant variables exceeds the number of observations, i.e. $N > T$. Moreover, the methods for which $\alpha \in (0, 1]$, from here on referred to as lasso-type estimators, perform subset selection by shrinking coefficient estimates to zero. They are potentially able to further improve forecasting performance by reducing the added variance of estimating parameters of irrelevant variables. The weights ω_j, $j = 1, \ldots, N$, allow for differential shrinkage on the parameters. Zou (2006) demonstrates that the use of cleverly chosen initial estimators as weights improves the selection performance by penalizing irrelevant variables to a higher degree than relevant variables. Common choices for initial estimators are the absolute values of OLS or ridge coefficients from a preceding estimation. Furthermore, it can be directly observed from (2.3) that the pre-determined set of relevant variables $\boldsymbol{w}_t$ is free of regularization and is therefore ensured to be included in the final model. Following Friedman et al. (2010), the solution to (2.3) can be efficiently obtained using a coordinate descent algorithm.

Whereas the earlier theory for the lasso has been developed in rather restrictive frameworks such as fixed designs (e.g. Knight and Fu, 2000; Zou, 2006), the properties of the lasso and its variants are becoming increasingly well understood in time series settings. One strand of time series related literature focusses on a framework with a fixed number of independent variables. This includes, among others, the work of Wang et al. (2007) who apply the (adaptive) lasso to models with autoregressive errors and derive estimation and selection consistency, and Yoon et al. (2013) who build on these results by estimating the autoregressive order directly from the data and by considering additional penalization methods. Hsu et al. (2008) derive the asymptotic theory for the lasso estimator under vector autoregressive (VAR) processes, and

Kock (2016) considers application of the lasso to both stationary and nonstationary autoregressive processes.

Others have explored the realm of double-asymptotics, allowing the number of candidate variables to grow along with the sample size. Nardi and Rinaldo (2011) consider the estimation of autoregressive (AR) models where the number of lags increase with the sample size. Song and Bickel (2011) consider the (group-)lasso to estimate VAR models where the number of candidate variables is allowed to increase, but the number of relevant variables is kept fixed. Kock and Callot (2015) also use the lasso for VAR estimation, while allowing the number of relevant variables to increase. They provide non-asymptotic bounds and sufficient conditions for asymptotic consistency of the predictions, parameter estimates and variable selection. Unfortunately the generality of their results comes at the cost of imposing independence and normality on the errors. Medeiros and Mendes (2016) show that the adaptive lasso estimator maintains its consistency under substantially weaker assumptions and that the estimates are asymptotically normal even under weakly dependent errors. These results hold for (conditionally) heteroskedastic processes as well, although efficiency gains can be made through the use of alternative weighting (e.g. Wagener and Dette, 2013; Ziel, 2016). Thus, research has progressed to a point where lasso-type estimators are theoretically justifiable in a stationary time series context and the applied econometrician is now required to choose between two appealing, though rather contrasting, approaches to modelling high-dimensional data.

Tuning

The implementation of lasso-type estimators requires the user to provide an a priori choice on the tuning parameters (λ, α). In the simulation exercises and the empirical application to follow, the tuning parameters are determined by obtaining the solution to (2.3) on a (100×1) grid of λ-values for the methods with a pre-determined α value or a (100×6) dimensional grid with (λ, α)-tuples for the (adaptive) elastic-net. We then use an information criterion (BIC or AIC) or time series cross-validation (CV) to select the optimal value(s). Time series CV is performed by reserving the first part of the sample to estimate the model under various settings of the tuning parameters after which the resulting models' fit are compared in a pseudo out-of-sample evaluation (Hyndman and Athanasopoulos, 2018). To illustrate, we adopt the threshold $c_T = \lceil \frac{2}{3} \times T \rceil$ and let $\boldsymbol{Z}_{c_T} = (\boldsymbol{W}_{c_T}, \boldsymbol{X}_{c_T})$, where $\boldsymbol{W}_{c_T} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{c_T})'$ and $\boldsymbol{X}_{c_T} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{c_T})'$. For a given value of the tuning parameter, say λ_j for $j = 1, \ldots, 100$, the model is estimated on $\boldsymbol{Z}_{c_T}$ to obtain the coefficient vector $\hat{\boldsymbol{\beta}}(\lambda_j)$. Next, a pseudo out-of-sample mean squared forecast error is calculated as $MSFE(\lambda_j) =$

$\frac{1}{T-c_T}\sum_{t=c_T+1}^{T}(y_{t+h} - \boldsymbol{z}_t'\hat{\boldsymbol{\beta}}(\lambda_j))^2$. This procedure is executed for all values of the tuning parameter in the predefined grid and the final tuning parameter is chosen as

$$\hat{\lambda} = \arg\min_{\lambda_j} \ \text{MSFE}(\lambda_j).$$

In time series settings, this method is often preferred over traditional k-fold CV, because the time structure of the data is kept intact.[2]

2.2.2 Factor models

The literature on factor models is vast, their use being motivated through the conceptualization of factors as unobserved and possibly dynamic processes related to the state of the economy that drive a large set of observed economic time series. Factor models attempt to summarize the candidate set $\boldsymbol{X}$ by a smaller number of factors and, in the dynamic case, their lagged realizations. In this factor framework, the variables in the candidate set admit the following representation

$$\boldsymbol{x}_t = \boldsymbol{\Lambda}(L)\boldsymbol{f}_t^* + \boldsymbol{e}_t, \tag{2.4}$$

where $\boldsymbol{\Lambda}(L) = (\boldsymbol{\lambda}_1(L), \ldots, \boldsymbol{\lambda}_N(L))'$, $\boldsymbol{\lambda}_i(L) = (\lambda_{i,1}(L), \ldots, \lambda_{i,s}(L))'$ and $\lambda_{i,j}(L)$ is a lag polynomial of possibly infinite order describing how variable i loads onto the dynamic factor j. The symbol $\boldsymbol{f}_t^*$ refers to an $(s \times 1)$ vector containing the common factors and $\boldsymbol{e}_t$ is a vector of idiosyncratic disturbances.

The majority of the literature on forecasting with factor models has, either explicitly or implicitly, relies on the assumption of finiteness of the lag polynomials $\lambda_{i,j}(L)$. This assumption allows the model to be cast in a static form with the representation

$$\boldsymbol{x}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{e}_t. \tag{2.5}$$

where $\boldsymbol{\Lambda}$ contains the coefficients in $\boldsymbol{\Lambda}(L)$, $\boldsymbol{f}_t = (\boldsymbol{f}_t^{*\prime}, \ldots, \boldsymbol{f}_{t-q}^{*\prime})'$ is a vector of size r with $s \leq r \leq (q+1)s$ and $\boldsymbol{e}_t = (e_{1t}, \ldots, e_{Nt})'$. The extension to the purpose of forecasting our generic model (2.1) follows naturally by substituting the candidate

[2] While standard k-fold CV is valid for purely autoregressive models with uncorrelated errors (Bergmeir et al., 2015), we observe time series CV to perform similarly in the simulations and superior in the empirical application.

variables for their factor representation:

$$\begin{aligned} y_{t+h} &= \boldsymbol{w}_t'\boldsymbol{\beta}_w + \boldsymbol{x}_t'\boldsymbol{\beta}_x + \epsilon_{t+h} \\ &= \boldsymbol{w}_t'\boldsymbol{\beta}_w + \boldsymbol{f}_t'\boldsymbol{\Lambda}'\boldsymbol{\beta}_x + \boldsymbol{e}_t'\boldsymbol{\beta}_x + \epsilon_{t+h} \\ &= \boldsymbol{w}_t'\boldsymbol{\beta}_w + \boldsymbol{f}_t'\boldsymbol{\beta}_f + u_{t+h}, \end{aligned} \tag{2.6}$$

with $\boldsymbol{\beta}_f = \boldsymbol{\Lambda}'\boldsymbol{\beta}_x$ and u_{t+h} being the composite error that includes the innovation ϵ_{t+h} and the loss of information from summarizing the data $\boldsymbol{e}_t'\boldsymbol{\beta}_x$. The reduction in dimension from N to r allows this model to be estimated with OLS and the dependent variable to be forecast as $\hat{y}_{T+h|T} = \boldsymbol{w}_T'\hat{\boldsymbol{\beta}}_w + \hat{\boldsymbol{f}}_T'\hat{\boldsymbol{\beta}}_{\hat{f}}$. Estimating the factors $\hat{\boldsymbol{f}}_T$ can be done with a wide variety of algorithms, the most common of which we discuss next.

The method of principal components (PC) is a popular means of extracting static factors. For any given k, which need not be equal to the true number of static factors r, the standard method of principal components (PC) obtains a $(T \times k)$ matrix of factor estimates and a $(N \times k)$ matrix of estimated loadings by solving the objective function

$$\left(\hat{\boldsymbol{\Lambda}}^k, \hat{\boldsymbol{F}}^k\right) = \underset{\boldsymbol{\Lambda}^k, \boldsymbol{F}^k}{\arg\min} \sum_t (\boldsymbol{x}_t - \boldsymbol{\Lambda}^k \boldsymbol{f}_t^k)' \boldsymbol{\Omega}^{-1} (\boldsymbol{x}_t - \boldsymbol{\Lambda}^k \boldsymbol{f}_t^k) \tag{2.7}$$

with $\boldsymbol{\Omega} = \boldsymbol{I}_N$ and subject to the normalization $\boldsymbol{\Lambda}^{k\prime}\boldsymbol{\Lambda}^k/N = \boldsymbol{I}_k$ and $\boldsymbol{F}^k = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T)'$ with $\boldsymbol{F}^{k\prime}\boldsymbol{F}^k$ being diagonal.

A drawback of forecasting with standard PC is that the quality of the estimated components that serve as inputs for the forecasting equation strongly depends on the structure inherent to the original data. For example, Boivin and Ng (2006) demonstrate that cross-sectional correlation in the idiosyncratic component of (2.5) is highly detrimental to the quality of the component estimates. In search for a more robust form of component estimation, they propose the use of weighted principal components (WPC) by replacing the unobserved inverted population covariance matrix $\boldsymbol{\Omega}^{-1}$ in (2.7) with a feasible estimate $\hat{\boldsymbol{\Omega}}^{-1}$. Boivin and Ng (2006, p. 185) propose several weighting rules to obtain feasible estimates such as their weighting 'rule SWa', where $\hat{\boldsymbol{\Omega}}^{-1}$ is diagonal with the i^{th} diagonal element equal to $\left(\frac{1}{T}\sum_{t=1}^{T}\hat{e}_{i,t}^2\right)^{-1}$. We explore the additional rules 'SWb', 'Rule1' and 'Rule2' proposed in their original paper as well and refer to them by their original names respectively.

Another cited disadvantage of principal component analysis is that every component is a linear combination of all variables, while a common empirical observation is that for any given component large groups of variables may carry small, non-zero

loadings (e.g. Stock and Watson, 2002b; Croux and Exterkate, 2011). Similar to the premise underlying the lasso, it may be favourable to estimate factors that depend only on a subset of the variables to reduce forecast variability and, when of interest, improve interpretability of the model. The solution brought forward in the literature takes the form of sparse principal component (SPC), variants of which occur in Jolliffe et al. (2003), Zou et al. (2006) and Shen and Huang (2008). More recently, Kristensen (2017) considers the use of SPC for macroeconomic forecasting and shows that, under suitable restrictions on the amount of shrinkage, the SPC estimator is consistent under assumptions similar to those in Stock and Watson (2002a). While no additional assumption on the sparseness of the loadings is required for its consistency, the use of SPC implicitly favours a sparse representation from the perspective of the classical bias/variance tradeoff. In this chapter we adopt the computationally beneficial approach of Shen and Huang to estimate the sparse principal components and refer the reader to their original paper for details.

An alternative method of imposing sparsity is proposed by Bai and Ng (2008a) who argue for forecasting with factor-augmented regressions by applying principal components to a subset of the predictors selected with the use of shrinkage estimators such as the lasso. Given the intuitive appeal of this approach and the documented improvement in performance by Bai and Ng, we include their $PC(LA)$-approach by applying the lasso for the purpose of subset selection in the first stage and extracting factors from that subset using standard PC in the second stage.[3]

Rather than casting the dynamic factor model (2.4) in the static framework (2.5), one may want to estimate the dynamic specification directly. Forni et al. (2000) propose a method to directly estimate (2.4) by obtaining the s dynamic factors on the basis of a consistent estimate of the population spectral density matrix. However, since the recovery of the dynamic factor relies on the estimation of a two-sided truncated filter, this approach does not work well for forecasting at the end of the sample. Accordingly, Forni et al. (2005a) propose an alternative approach that decomposes the long run variance of the candidate set into contributions by the common and idiosyncratic components and estimates the factor loadings such that the share of the long rung variance attributable to the common component is maximized. This method is henceforth referred to as FHLR (Forni, Hallin, Lippi and Reichlin).

[3]Others have also considered the reverse order, i.e. first extracting principal components from the data and then performing shrinkage on those components (e.g. Stock and Watson, 2012; Kim and Swanson, 2014). Yet another possibility is to apply shrinkage alongside factor estimation by sparsely estimating the idiosyncratic component (e.g. Luciani, 2014; Hansen and Liao, 2019). These approaches, however, are not pursued here as they are less related to the central questions examined in this chapter and since their theoretical properties and empirical performance are well documented in the cited papers.

An alternative approach of explicitly modelling the dynamics in a factor model is to explicitly incorporate them into a likelihood function. The idea of estimating static factors by maximum likelihood dates back to the early work of Chamberlain and Rothschild (1983). More recently, however, Doz et al. (2011) and Doz et al. (2012) derive the theory for maximum likelihood estimation of factor models under much less restrictive assumptions on the dynamic structure of the factors and the idiosyncratic component. While their model estimation procedure relies on the use of the Kalman filter and a relatively strict set of assumptions, such as a diagonal covariance matrix of the idiosyncratic component,Doz et al. show that certain deviations away from these assumptions are asymptotically negligible, thereby justifying the method for a much broader class of data generating processes. We incorporate the maximum likelihood procedure in Doz et al. (2012) and will henceforth refer to this method as DGR (Doz, Giannone and Reichlin).

Finally, in recent contributions Forni et al. (2015, 2018) develop a method to obtain estimates of the dynamic components without imposing finiteness on the factor space. Under general assumptions, the authors derive one-sided representation of the dynamic factor model that can be estimated and used for forecasting. Throughout the chapter we will refer to this method of forecasting as FHLZ (Forni, Hallin, Lippi and Zaffaroni), while referring the interested reader to the cited papers for details.

Tuning

All of the methods described above require an a priori choice for the number of factors. As such, much attention has been given to the development of data driven criteria that may aid the researcher in this choice absent of knowledge of the true number of factors. The reference criteria for static factor models in most contributions are those provided by Bai and Ng (2002), who propose two classes of information criteria that minimize the variance of the idiosyncratic component subject to a penalty depending on both N and T. This method, however, is often documented to overestimate or underestimate the true number of factors (e.g. Forni et al., 2009), on the grounds of which we employ several alternative criteria in the comparisons to follow. We consider methods that use the same type of information criteria with an extra tuning parameter (Alessi et al., 2010) or that directly exploit the structure of the eigenvalues in the sample covariance matrix (Onatski, 2010; Ahn and Horenstein, 2013). For the dynamic factor models we employ the criteria of Hallin and Liška (2007) to select the number of dynamic components s. The DGR approach requires specification of the autoregressive order of the dynamic factors. This is determined by obtaining initial estimates of the factors by principal components and fitting a VAR model on these

estimates with the lag order being selected by the AIC. Finally, we implement the FHLZ method by randomly dividing the cross section of N time series in $\lfloor \frac{N}{q+1} \rfloor$ blocks on which we: (1) estimate VARs with their order determined by the AIC, (2) recover the dynamic components and (3) use these dynamic components and their lags to predict the dependent variable by an OLS projection.[4] This three-step process is repeated 50 times and the predictions are averaged over all iterations to remove the added noise from the cross-sectional sampling.

In the remainder of the chapter we will stick to the convention of tabulating results only for the tuning method that obtains the best performance on the factor model under consideration. Additional comments on the performance of other tuning methods are provided whenever deemed informative.

2.3 Simulation study

Our simulation study can broadly be categorized into three main sections, namely simulations on a DGP with (1) stationary observable variables with a sparse coefficient vector, (2) stationary common factors driving a large set of time series, and (3) non-stationary and cointegrated variables. In every category, we vary additional DGP characteristics such as the level of non-sphericity in the error, the number of common factors and the strength of the cointegration relationship.

Stationary observable variables

We generate the first set of DGPs as stationary processes where the dependent variable depends on five observable explanatory variables and a possibly autoregressive error term:

$$\begin{aligned} y_{t+1} &= \boldsymbol{x}_t' \boldsymbol{\beta}_x + \sqrt{\theta}\epsilon_{t+1} \\ (1-\alpha L)\epsilon_{t+1} &= v_{t+1} \end{aligned} \tag{2.8}$$

with $\boldsymbol{x}_t \sim \mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma}_N)$ and $v_{t+1} \sim \mathbb{N}(0,1)$. Let $\boldsymbol{\iota}_5$ be a (5×1) vector of ones and $\mathbf{0}_{N-5}$ an $((N-5)\times 1)$ vector of zeros, then $\boldsymbol{\beta}_x = (\boldsymbol{\iota}_5', \mathbf{0}_{N-5}')'$. The population covariance matrix takes on the form $\boldsymbol{\Sigma}_N = (\sigma_{i,j})_{i,j=1}^N$ with $\sigma_{i,j} = \rho^{|i-j|}$. Hence, $\boldsymbol{\Sigma}_N$ is a Toeplitz-matrix that allows for regulation of the degree of pairwise correlation between variable i and j by varying the single parameter ρ. In addition, we randomize the cross-sectional

[4] To take into account the complete dynamic structure, predictions ought to be obtained by filtering the estimated factors as in Forni et al. (2018). However, we find that the direct OLS projection frequently outperforms the filtered predictions, especially for multi-step predictions in the empirical application, which motivates our choice of implementation.

order of the newly generated variables prior to the construction of $\boldsymbol{y}$ in order to avoid a clustering of correlation in neighbouring variables. Furthermore, the signal-to-noise ratio is controlled by setting $\theta = \frac{1-\alpha^2}{10}\boldsymbol{\beta}_x'\boldsymbol{\Sigma}_N\boldsymbol{\beta}_x$, which keeps the population signal-to-noise ratio constant for changes in dimensionality of the model, as well as changes in the degree of serial correlation.

At every trial we generate $T = 100$ observations to which we apply all of the methods covered in section 2.3. For the shrinkage estimators we generate the 1-step ahead forecast as $\hat{y}_{T+1|T} = \boldsymbol{x}_T'\hat{\boldsymbol{\beta}}_x$, whereas the predictions from factor models are obtained as $\hat{y}_{T+1|T} = \hat{\boldsymbol{f}}_T'\hat{\boldsymbol{\beta}}_F$. This procedure is repeated over $J = 1,000$ trials and we evaluate the forecast performance of model i by the mean squared forecast error (MSFE)

$$\text{MSFE}_i = \frac{1}{J}\sum_{j=1}^{J}(y_{j,T+1} - \hat{y}^i_{j,T+1|T})^2. \tag{2.9}$$

The MSFE is reported relative to the MSFE of the optimal, though infeasible, OLS oracle method which forecasts the dependent variable by applying OLS to the five relevant variables only. As a measure of the estimation accuracy we calculate the mean squared error as

$$\text{MSE}_i = \frac{1}{J}\sum_{j=1}^{J}\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^i_j\right\|_2^2, \tag{2.10}$$

and, again, report the MSE relative to the OLS oracle procedure. Given the misspecified nature of the factor models on the current set of DGPs, this metric is reported for the shrinkage estimators only.

The selection performance of the shrinkage estimators is evaluated according to two standard metrics; the metric *consistent* depicts the fraction of trials in which the shrinkage estimators exactly identify the sparsity pattern by selecting the five relevant variables only, whereas *conservative* depicts the fraction of trials in which at least all five relevant variables are included. Finally, we also report the average number of variables included by each method as *#variables*. Detailed results regarding the shrinkage estimators are gathered in Table 2.1 - 2.2. The performance of the factor models is tabulated in Table 2.3.

The results in Table 2.1 emphasize the effect of changes in dimensionality by leaving out any cross-sectional and serial correlation ($\rho = \alpha = 0$). Panel A reports results for the low-dimensional case ($N = 10$). In terms of the mean squared forecast

Table 2.1 Stationary observed variables: the effect of dimensionality

	OLS	ridge		las		adalas		en		adaen	
		BIC	CV	BIC	CV	BIC	CV	BIC	CV	BIC	CV
Panel A: $N = 10$											
RMSFE	1.05	1.11	1.13	1.08	1.08	**1.01**	1.05	1.08	1.08	1.01	1.05
RMSE	2.13	2.47	2.91	2.07	2.35	1.21	1.84	2.07	2.46	1.21	1.95
consistent	0%	0%	0%	27%	13%	84%	52%	27%	11%	84%	35%
conservative	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
#variables	10.00	10.00	10.00	6.45	7.91	5.21	6.10	6.45	8.10	5.21	6.89
Panel B: $N = 50$											
RMSFE	1.92	1.75	1.85	1.20	1.20	**1.04**	1.12	1.20	1.21	1.04	1.13
RMSE	19.09	16.15	17.91	5.05	4.74	1.65	3.42	5.06	4.81	1.65	3.95
consistent	0%	0%	0%	12%	3%	60%	23%	12%	3%	60%	15%
conservative	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
#variables	50.00	50.00	50.00	8.31	15.69	5.85	11.98	8.32	15.82	5.85	16.42
Panel C: $N = 100$											
RMSFE	-	-	7.78	1.28	1.24	**1.08**	1.09	1.28	1.24	1.08	1.10
RMSE	-	-	139.42	6.85	5.90	2.69	3.01	6.85	5.96	2.67	3.25
consistent	-	-	0%	8%	3%	33%	15%	8%	3%	33%	12%
conservative	-	-	100%	100%	100%	100%	100%	100%	100%	100%	100%
#variables	-	-	100.00	9.75	19.47	6.56	10.51	9.76	19.70	6.58	11.04

Notes: Numerical entries in this table are averages obtained over 1,000 simulations relative to the OLS oracle method for all evaluation metrics described in section 2.3. Results are given for the low, mid and high-dimensional case in panel A,B and C respectively.

error penalized regression performs at least as well as OLS, with the exception of ridge regression. The latter is unsurprising given that ridge regression does not impose sparsity and is a biased estimator that aims to reduce the MSE through a favourable bias-variance trade-off. The ability to do so, however, hinges on the presence of multi-collinearity, which is not an issue in the current set-up. Focussing on the lasso-type methods, we observe that the forecast performance of the adaptively weighted variants is superior to their non-weighted counterparts and, with RMSFEs of 1.01, is comparable to the infeasible oracle estimator. Concerning the selection performance, three results stand out. First, selection of the tuning parameter(s) by the BIC seems to lead more frequently to exact identification of the five relevant explanatory variables compared to cross-validation. Second, an adaptive weighting of the tuning parameter substantially improves the consistent selection scores and results in smaller models on average. Third, all methods considered are able to include the five relevant variables in all trials.

While promising, the results so far are derived in a low-dimensional setting where the gain relative to traditional OLS is small and the often cited 'curse of dimensionality' is far from an issue. Accordingly, panel B-C display the performance for

Table 2.2 Stationary observed variables: the effect of correlation

		OLS	ridge		las		adaLas		en		adaen	
ρ	α		BIC	CV	BIC	CV	BIC	CV	BIC	CV	BIC	CV
Panel A: RMSFE												
0.0	0.0	1.92	1.75	1.85	1.20	1.20	**1.04**	1.12	1.20	1.21	1.04	1.13
0.6	0.0	1.94	1.52	1.56	1.12	1.16	1.02	1.12	1.12	1.18	**1.02**	1.14
0.6	0.6	1.88	1.49	1.51	1.13	1.14	1.03	1.09	1.13	1.15	**1.03**	1.11
Panel B: Consistent												
0.0	0.0	0%	0%	0%	12%	3%	60%	23%	12%	3%	60%	15%
0.6	0.0	0%	0%	0%	4%	2%	44%	16%	4%	2%	44%	11%
0.6	0.6	0%	0%	0%	4%	2%	48%	16%	4%	2%	48%	11%
Panel C: # variables												
0.0	0.0	50.00	50.00	50.00	8.31	15.69	5.85	11.98	8.32	15.82	5.85	16.42
0.6	0.0	50.00	50.00	50.00	9.28	15.45	6.24	11.49	9.30	16.22	6.26	15.48
0.6	0.6	50.00	50.00	50.00	9.20	15.63	6.16	11.55	9.20	16.40	6.17	16.15

Notes: see notes in 2.1. The metrics considered are: (A) the RMSFE , (B) Consistent, and (C) the number of variables. Within each panel the different rows correspond to different settings of the degree of cross-sectional correlation (ρ) and serial correlation (α).

$N = 50$ and $N = 100$. The relative forecasting performance of OLS and ridge regression deteriorates and the difference in RMSFE with the sparsity inducing methods becomes more pronounced, despite the unreported MSFEs of the latter methods increasing along with the dimensionality as well. The detrimental effects of an increase in dimensionality are perhaps most apparent in the selection performance, with exact identification of the sparsity pattern occurring at substantially lower frequencies. Given that the conservative selection remains 100%, the drop in consistent selection necessarily stems from the inclusion of additional irrelevant variables, most likely due to randomly induced collinearity. Indeed, the increase in the number of variables selected in the higher dimensional settings supports this conjecture.

A well-known problem for the lasso is the presence of multi-collinearity in the data, especially between relevant and irrelevant variables, which can lead to inconsistencies in the selection of the correct variables (e.g. Zhao and Yu, 2006; Zou, 2006). As such, we examine the forecasting and selection performance under varying degrees of cross-sectional and serial correlation in Table 2.2, whilst keeping the dimension fixed at $N = 50$. Noteworthy is that while the MSFE increases for all methods when introducing a higher degree of cross-sectional correlation (unreported), the relative MSFE decreases for ridge regression and varies only marginally for the lasso-based regressions. The former finding is in line with the proclaimed benefits of L_2-penalization under multi-collinearity, whereas the latter finding hints that the presence of cross-sectional correlation does not seem to affect the forecasting performance of lasso-type estimators more than OLS. Panel B clearly depicts the deterioration in selection per-

Table 2.3 Stationary observed variables: factor models

	PC	WPC				SPC	PC(LA)	FHLR	FHLZ	DGR
		SWa	SWb	Rule1	Rule2					
Panel A: $N = 50, \rho = 0$										
RMSFE	**9.06**	9.44	9.17	9.85	9.85	9.10	9.16	9.82	9.75	9.68
nvar	3.40	1.92	2.48	1.00	1.01	3.40	3.30	1.00	1.00	1.00
Panel B: $N = 50, \rho = 0.6$										
RMSFE	**2.57**	2.69	2.67	3.24	4.17	2.59	3.39	4.66	4.79	4.68
nvar	10.00	9.79	9.96	7.17	4.89	9.98	5.16	1.00	1.00	1.00

Notes: see notes in 2.1. Panel A lists results for a DGP with uncorrelated variables, whereas panel B lists results for a DGP allowing for a maximum population correlation of 0.6 between variables.

formance after the introduction of cross-sectional correlation. While the unreported metric for conservative selection remains 100% for all methods, the consistent selection is strongly affected by the presence of cross-sectional correlation. In line with the aforementioned reasoning on the selection performance in high-dimensional settings, this implies that high levels of collinearity lead to larger models with irrelevant variables being erroneously included at higher frequencies. Finally, the method by which we scale the idiosyncratic noise term controls for the increased variance induced by serial correlation and, consequently, the introduction of serial correlation has little effect on the relative forecasting or selection performance.

Finally, in Table 3 we examine the predictive capabilities of factor models in the current framework. For each factor model, the results are reported for the factor selection method that delivers the best performance. Unsurprisingly, on a DGP absent of common components the factor models display inferior performance compared to the shrinkage estimators in Table 2.2. While the forecast accuracy worsens less when the variables in the dataset are correlated (Panel B) and when the information criterion selects a higher number of components, failure to include as many components as there are variables in the original dataset inevitably leads to a loss of information that negatively affects the forecasting performance. As a result, the *PC*-type criteria of Bai and Ng (2002) tend to deliver the best forecast accuracy here as they select more components on average. On the contrary, the dynamic factor models demonstrate relatively poor performance mainly as a result of the Hallin and Liška criterion selecting only a single dynamic factor in all simulation trials.

Stationary common factors

We next turn to the case where a small number of common factors drive a larger set of time series. The data-generating process contains an approximate factor structure

and is a simplified version of the Stock and Watson (2002a) set-up recently employed by Kristensen (2017):

$$\begin{aligned} x_{it} &= \boldsymbol{\lambda}_i' \boldsymbol{f}_t + e_{it} \\ (1 - \alpha L) e_{it} &= (1 + \theta^2) v_{it} + \theta v_{i+1,t} + \theta v_{i-1,t} \end{aligned} \tag{2.11}$$

with $\boldsymbol{\lambda}_i$, $\boldsymbol{f}_t \overset{iid}{\sim} \mathbb{N}(\mathbf{0}, \boldsymbol{I}_r)$. The random variable $v_{i,t}$ drives the idiosyncratic component and is generated from a standard normal distribution. We impose sparsity in the loadings by setting a fraction τ of them equal to zero. While sparsity here simply refers to the presence of exact zero elements in the loadings, our approach of setting a fraction of all loadings equal to zero does not contradict the classic assumption of dense factor loadings, i.e. $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}/N \to \boldsymbol{I}_r$. As a result, even though the method of sparse principal components is expected to be more efficient here, the use of 'non-sparse' factor models remains theoretically justifiable. The variable to forecast is generated as

$$y_t = \boldsymbol{f}_t' \boldsymbol{\beta}_f + \epsilon_t \tag{2.12}$$

where $\boldsymbol{\beta}_f$ is an $(r \times 1)$ vector of ones and ϵ_t is a standard normal error term. Recall that the shrinkage estimators attempt to forecast y_{T+1} as $\hat{y}_{T+1|T} = \boldsymbol{x}_t' \hat{\boldsymbol{\beta}}_x$, whereas the factor models use the extracted factors to construct the forecast $\hat{y}_{T+1|T} = \hat{\boldsymbol{f}}_t' \hat{\boldsymbol{\beta}}_{\hat{f}}$. Forecasting performance is measured on the basis of the MSFE relative to the factor-augmented regressions with the true number of factors, calculated by standard PC. The two-step procedure calls for an additional metric measuring the estimation precision of the factor estimates in the first step. Following Doz et al. (2012) and Kristensen (2017), we report the trace R^2 as a measure to determine how well the estimated factors span the space of the true factors, calculated as

$$R_F^2 = \frac{Tr\left(\boldsymbol{F}'\hat{\boldsymbol{F}}(\hat{\boldsymbol{F}}'\hat{\boldsymbol{F}})^{-1}\hat{\boldsymbol{F}}'\boldsymbol{F}\right)}{Tr\left(\boldsymbol{F}'\boldsymbol{F}\right)}, \tag{2.13}$$

where $\hat{\boldsymbol{F}} = (\hat{\boldsymbol{f}}_1, \ldots, \hat{\boldsymbol{f}}_T)'$ and $Tr(\cdot)$ represents the trace function. While the shrinkage estimators obviously do not extract factors on the observed variables, the trace R^2 remains informative when interpreted as a measure of the accuracy with which the factor space is approximated by the subset of variables chosen by a given shrinkage estimator. Hence, for the shrinkage estimators we estimate

$$R_X^2 = \frac{Tr\left(\boldsymbol{F}'\boldsymbol{X}_S(\boldsymbol{X}_S'\boldsymbol{X}_S)^{-1}\boldsymbol{X}_S'\boldsymbol{F}\right)}{Tr\left(\boldsymbol{F}'\boldsymbol{F}\right)}, \tag{2.14}$$

where $\boldsymbol{X}_S$ denotes the subset of variables included by the method under consideration. The results for the set of DGPs with a single factor driving the time series are reported in Table 2.4 and for the case of four common factors in Table 2.5. To focus the comparison on differences between the factor extraction methods, rather than the factor selection methods, we report the results using the true number of factors only.[5]

Table 2.4 - panel A reveals that the the factor models manage to slightly outperform the shrinkage estimators on a DGP where the population covariance matrix of the idiosyncratic component is diagonal, i.e. $\alpha = 0$ and $\theta = 0$. The trace R^2s are close to unity, which for the factor models implies accurate recovery of a rotation of the unobserved factor. For the shrinkage estimators, the high R^2s indicate that the limited number of variables chosen seems to be sufficient for a reasonable approximation of the factor space. This finding is in accordance with the proposition of De Mol et al. (2008) who reason that the factor-induced collinearity in the candidate set allows for a few appropriately selected variables to capture the majority of the covariance in the data and to span approximately the same space as the common factors. Finally, ridge regression performs slightly worse than the lasso-type estimators and the OLS estimator displays the lowest forecast accuracy of all methods, despite obtaining the highest R^2. This illustrates that on the kind of non-sparse DGPs here considered, in which each individual variable possesses only little explanatory power over the dependent variable of interest, the application of shrinkage reduces model complexity by favouring the effect of those variables with high predictive power. Our results demonstrate that in such cases, the forecasting performance can benefit from a favourable bias-variance trade-off.

According to De Mol et al. (2008), forecasts from lasso-type estimators should not be expected to outperform correctly specified factor-augmented regressions, since the subset of the data proposed by methods employing an L_1-penalty offers merely an approximation to the factor space and variable selection under high degrees of collinearity is known to be unstable. Indeed, panel B of Table 2.4 shows that the shrinkage estimators still underperform the factor models even when the component loadings are sparse. However, in panel C we observe that, after the introduction of substantial non-sphericity in the idiosyncratic component, the forecasting performance is tilted in favour of the shrinkage estimators. Under high levels of non-sphericity the factor models have difficulty in accurately estimating the unobserved factors, as indicated

[5] While the performance differentials between factor extraction methods remain qualitatively similar under the use of factor selection criteria, we do note the general finding that under strong forms of non-sphericity and a DGP with four latent factors all criteria tend to understimate the true number of factors, with the exception of the *PC*-type criteria which heavily overestimate the true number of factors. All factor selection methods are more accurate under spherical idiosyncratic disturbances.

Table 2.4 DGP with one common factor

	PC	WPC				SPC	PC(LA)	FHLR	FHLZ	DGR	las	adalas	en	adaen	ridge	ols
		SWa	SWb	Rule1	Rule2											
Panel A: $\alpha/\theta/\tau = 0/0/0$																
RMSFE	1.00	0.98	0.96	1.12	1.36	1.00	1.09	0.96	1.03	**0.96**	1.15	1.17	1.09	1.07	1.35	1.87
nvar	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	20.14	15.81	37.55	37.92	50.00	50.00
R^2	0.96	0.97	0.97	0.95	0.92	0.96	0.96	0.97	0.96	0.97	0.98	0.98	0.99	0.99	0.99	0.99
Panel B: $\alpha/\theta/\tau = 0/0/0.4$																
RMSFE	1.00	0.95	0.93	1.18	1.54	0.98	1.03	0.92	1.03	**0.92**	1.11	1.07	1.11	1.06	1.38	1.80
nvar	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	16.96	14.45	27.08	36.29	50.00	50.00
R^2	0.94	0.95	0.95	0.92	0.87	0.94	0.94	0.95	0.93	0.95	0.97	0.96	0.97	0.98	0.98	0.98
Panel B: $\alpha/\theta/\tau = 0.5/1/0.4$																
RMSFE	1.00	0.97	0.98	1.00	1.06	0.98	1.04	0.95	0.80	0.96	0.26	**0.26**	0.26	0.26	0.27	0.30
nvar	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	39.35	33.23	39.42	33.19	50.00	50.00
R^2	0.41	0.41	0.42	0.42	0.39	0.42	0.40	0.43	0.55	0.44	1.00	0.99	1.00	0.99	1.00	1.00

Notes: The reported RMSFEs are relative to the PC estimator that uses a single components in the forecasting equation. Each panel corresponds to a different setting of the degree of serial correlation (α), cross-sectional correlation (θ) and sparsity in the loadings (τ).

Table 2.5 DGP with four common factors

	PC	WPC				SPC	PC(LA)	FHLR	FHLZ	DGR	las	adalas	en	adaen	ridge	ols
		SWa	SWb	Rule1	Rule2											
Panel A: $\alpha/\theta/\tau = 0/0/0$																
RMSFE	1.00	1.04	0.96	1.23	1.63	1.00	1.11	0.96	1.22	**0.96**	1.22	1.20	1.16	1.13	1.24	1.88
nvar	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	19.89	16.17	38.44	39.83	50.00	50.00
R^2	0.96	0.96	0.97	0.95	0.90	0.96	0.93	0.97	0.93	0.97	0.93	0.91	0.97	0.97	0.99	0.99
Panel B: $\alpha/\theta/\tau = 0/0/0.4$																
RMSFE	1.00	0.94	0.92	1.23	1.90	1.00	1.07	0.93	1.13	**0.92**	1.17	1.15	1.15	1.11	1.24	1.69
nvar	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	22.26	18.15	36.56	39.96	50.00	50.00
R^2	0.94	0.95	0.95	0.91	0.81	0.94	0.90	0.95	0.91	0.95	0.93	0.92	0.96	0.96	0.98	0.98
Panel B: $\alpha/\theta/\tau = 0.5/1/0.4$																
RMSFE	1.00	0.98	1.00	1.01	1.16	1.00	0.98	0.97	0.84	0.97	0.33	0.33	0.33	**0.33**	0.33	0.36
nvar	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	43.30	37.92	43.26	37.90	50.00	50.00
R^2	0.51	0.51	0.51	0.47	0.41	0.50	0.48	0.52	0.55	0.52	0.99	0.97	0.99	0.97	1.00	1.00

Notes: See notes in Table 4. The RMSFE is relative to the standard PC estimator that extracts four components.

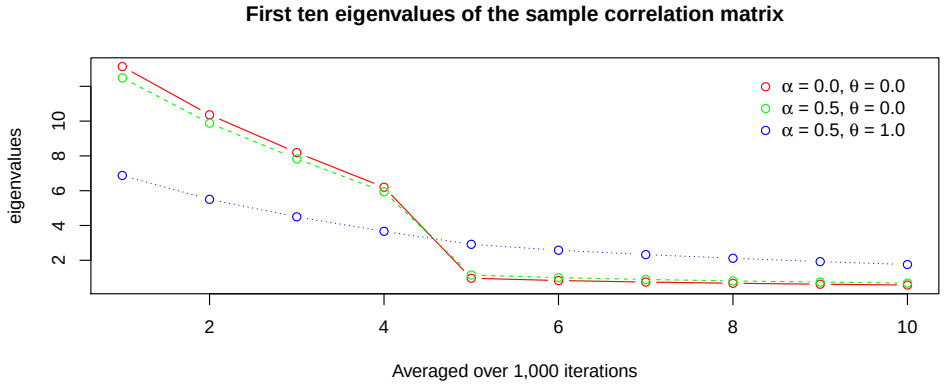

Figure 2.1: Visualization of the explanatory power of the first ten common components.

by the decrease in trace R^2s, whereas the shrinkage estimators tend to select a higher number of variables on average and, as a result, are able to maintain accurate approximation of the factor space. These patterns are similarly observed in the DGP with four factors, the results of which are displayed in Table 2.5, and provide a clear argument in favour of lasso-type estimation on data possessing factor structures with potentially non-spherical idiosyncratic components.

Upon further analysis, the introduction of cross-sectional correlation in the error term in (2.11) appears to be the main culprit for the deterioration in factor quality estimates. In the DGP with four factors, the percentage of the variance in the candidate set $\boldsymbol{X}$ explained by the first four standard estimated principal components is 72.3% before the introduction of cross-sectional correlation ($\alpha = 0.5$, $\theta = 0$) and 41.1% afterwards ($\alpha = 0.5$, $\theta = 1$). This is visualized in Figure 2.1, where we display the ten largest eigenvalues of the sample correlation matrix corresponding to the first ten principal components. We conjecture that the correlation between the series in the candidate set that is induced by the idiosyncratic component obscures the factor-induced variation, thereby reducing the precision by which the factors are estimated. Apparently, the large number of non-zero off-diagonal parameters in the covariance matrix of the errors cannot simply be ignored, or estimated accurately enough by the weighted principal component estimators, while maintaining precise estimates of the underlying factors.

Non-stationary and cointegrated variables

The presence and consequences of non-stationary predictors in regression frameworks are well-understood and numerous tests and solutions have been proposed to correct for non-stationarity. Accordingly, in the majority of simulations and empirical work the implicit assumption is maintained that the researcher is able to successfully identify non-stationarity and all variables found to be integrated of order one or higher are transformed to stationarity by taking appropriate differences. However, situations are frequently encountered where the order of integration remains ambiguous (e.g. fractionally integrated variables or weakly cointegrated variables). In addition, the act of "correcting" for non-stationarity by differencing the variables comes at the cost of losing information captured in the levels of the variables. The literature on cointegration shows that long-run relationship between non-stationary variables can exist, relationships that are impossible to recover when using differenced variables. Here we examine the potential of lasso-type estimators in identifying and utilizing cointegrating relationships for forecasting in high-dimensional systems.

The potential for penalized regression in recognizing cointegrating relationships has recently been explored by Wilms and Croux (2016), Liao and Phillips (2015) and Liang and Schienle (2019) who all consider the use of penalized regression in automated vector error correction model estimation. These novel and insightful contributions, however, require a non-standard and fairly technical implementation. In an attempt to avoid placing this burden on the researcher, we focus on the use of an intuitive single equation model rather than a multivariate model. An investigation of regularized VECM estimation is postponed to Chapter 5.

We generate the data as an error correction model:

$$\begin{aligned} \Delta y_t &= \alpha \left(y_{t-1} - \sum_{i=1}^{3} \beta_i x_{i,t-1} \right) + \epsilon_{j,t} \\ x_{i,t} &= x_{i,t-1} + \epsilon_{j+1,t} \qquad\qquad i = 1,2,3,\ j = 1,2,3 \end{aligned} \tag{2.15}$$

where the stationarity condition is given by $-2 < \alpha < 0$ and $\boldsymbol{\epsilon}_t \sim \mathbb{N}(\mathbf{0}, \boldsymbol{I}_4)$. In addition to the three variables $x_{i,t}$ for $i = 1, \ldots, 3$ that cointegrate with y_t we add a number of irrelevant variables to the candidate set $\boldsymbol{X}$. The high sample correlations induced by variables that are integrated of order one, i.e. I(1), may have adverse consequences on the prediction and selection performance of the shrinkage estimators. Accordingly, we perform two sets of simulations; one in which the irrelevant variables are generated according to (2.8) with $\rho = 0.5$, $\alpha = 0$, and one in which half of

the irrelevant variables are generated similarly, but the other half are generated as random walks, i.e. $\Delta x_{k,t} = \epsilon_{k,t}$ with $\epsilon_{k,t} \sim \mathbb{N}(0,1)$. The two sets of simulations are simply referred to as "Stationary" and "Non-Stationary". As an example, for a candidate set $\boldsymbol{X}$ of size $N = 50$ that is generated in the Non-stationary set, the first three variables will be I(1) but cointegrated with the dependent variable. In the set of irrelevant variables, $\lceil \frac{N-3}{2} \rceil = 24$ are I(0) and $\lfloor \frac{N-3}{2} \rfloor = 23$ are I(1). In congruence with the preceding simulations, we generate 1,000 one-step ahead forecasts and report the metrics RMSFE and RMSE relative to the oracle OLS procedure as measures of prediction and selection performance respectively. The selection performance is, again, measured with the metrics *consistent*, *conservative* and *#variables*. The use of factor models is excluded from this section on the grounds that extracted factors can contain linear combinations of non-stationary variables and, hence, will be integrated of order one. Indeed, the presence of stochastic trends in the factors necessitates the use of alternative methods, such as the factor-augmented error correction model by Banerjee and Marcellino (2009), the forecasting performance of which is considered in Banerjee et al. (2014a), or estimation of the factors in a VECM framework in the spirit of Barigozzi et al. (2016a,b). While these methods are excluded from the analysis here, they are considered in detail in Chapter 5.

We present the main results for the remaining estimators in Table 2.6, where the adjustment rate is fixed at $\alpha = -1$ and all tuning parameters are optimized based on the BIC. The effect of changes in the adjustment rate are further explored in Table 2.7.

Focussing on the predictive capabilities first, the RMSFEs in panel A of Table 2.6 demonstrate a superior performance of the L_1 methods. The minimum RMSFE, denoted in bold, is always obtained by an adaptively weighted lasso-type estimator. Notwithstanding an overall decrease in forecasting performance relative to the OLS oracle procedure, the comparative advantage of lasso-type methods relative to OLS or ridge becomes more pronounced for higher dimensions. The advantage of adaptive weighting over non-weighted estimation is substantial for the dimensions $N = 10$ and $N = 50$, but seems to diminish at $N = 100$. This most likely results from a deterioration in quality of the initial estimator, thereby highlighting the importance of finding good initial estimators in the high-dimensional setting.[6] The estimation accuracy of the cointegrating vector, as measured by the RMSE, follows the same pattern as the prediction performance, with adaptively weighted estimation providing the highest accuracy and outperforming OLS even in the low-dimensional setting.

[6] This issue is particularly prominent in the theoretical analysis presented in Chapter 4.

Table 2.6 Cointegrated variables

	Stationary			Non-Stationary		
	N=10	N=50	N=100	N=10	N=50	N=100
Panel A: RMSFE						
OLS	1.10	1.83	-	1.11	2.20	-
ridge	1.37	2.10	18.84	1.40	1.74	6.88
las	1.17	1.51	1.74	1.17	1.58	1.82
adalas	**1.03**	**1.09**	1.45	**1.05**	1.34	**1.60**
en	1.17	1.51	1.74	1.18	1.58	1.81
adaen	1.03	1.09	**1.43**	1.05	**1.34**	1.63
Panel B: RMSE						
OLS	9.38	106.70	-	7.48	89.98	-
ridge	9.89	64.72	46.26	11.61	51.82	46.61
las	4.22	8.21	10.64	5.31	18.88	26.90
adalas	2.16	3.25	8.37	2.51	16.39	24.86
en	4.22	8.20	10.78	5.33	18.98	27.10
adaen	2.16	3.24	8.08	2.52	16.46	25.14
Panel C: Consistent						
las	29.9%	20.1%	18.2%	9.8%	0.2%	0.0%
adalas	81.6%	62.4%	33.8%	63.8%	4.4%	0.2%
en	29.9%	20.0%	18.1%	9.9%	0.2%	0.0%
adaen	81.2%	62.2%	33.5%	63.6%	4.1%	0.2%
Panel D: Conservative						
las	99.5%	93.1%	88.5%	99.6%	82.5%	64.1%
adalas	99.8%	99.6%	91.2%	99.9%	79.3%	58.8%
en	99.5%	93.2%	88.5%	99.6%	82.3%	63.8%
adaen	99.8%	99.6%	91.6%	99.9%	79.3%	58.2%
Panel E: #Variables						
las	4.53	6.29	6.65	5.35	9.97	12.17
adalas	3.24	3.75	5.71	3.49	7.59	10.17
en	4.53	6.30	6.72	5.35	9.97	12.23
adaen	3.24	3.75	5.66	3.49	7.61	10.13

Notes: Numerical entries in this table are averages obtained over 1,000 simulations relative to the OLS oracle estimator that estimates the cointegrating vector with the cointegrated variables only. The methods considered are listed in the first column, whereas the evaluation metrics are divided across panels A-E. The results under 'Stationary' are derived on a DGP absent of irrelevant I(1) variables, whereas those listed under 'Non-Stationary' are derived on DGPs that do contain irrelevant I(1) variables.

The selection performance is depicted in the remaining three panels of Table 2.6. Panel C depicts the fraction of trials in which the lasso-type methods identify the sparse cointegrating relationship exactly. Again, the adaptively weighted variants show superior performance. Exact identification, however, occurs at considerably lower rates in higher dimensional settings, with the decline in selection performance being most notable for the adaptively weighted estimators. A direct comparison between the scores for the consistent metric obtained on the stationary and non-stationary sets reveals that the presence of irrelevant I(1) variables negatively affects the selection performance. We conjecture that the inevitable high correlation between the non-stationary variables in levels, regardless of their relevance to the dependent variable, increases the difficulty in identifying the correct subset. Given that exact identification seems to be overly ambitious in this framework, we turn our attention to conservative selection. Absent of irrelevant non-stationary variables in the candidate set, the lasso-type methods almost always include at least all relevant variables. With the inclusion of additional I(1) variables, we observe a worsening of the conservative selection, especially at higher dimensions, albeit not to levels as inadequate as observed for the *consistent* selection. Finally, the reason for *conservative* selection staying at reasonable levels can at least partly be attributed to the growing model size along increases in dimensionality. More irrelevant variables tend to be included when estimating on a larger candidate set and this effect is particularly apparent when non-stationary variables are present. Despite the faulty model selection characteristics in this non-stationary framework, the reduction in variance by excluding at least part of the irrelevant variables contributes enough to obtain a superior forecasting performance. Hence, for the applied researcher whose main interest lies in forecasting rather than model interpretation this somewhat naive application of lasso-type methods to cointegrated data in levels delivers substantial benefit.

The results so far are based on the somewhat idealized adjustment rate of $\alpha = -1$. If the adjustment rate would be closer to the lower boundary of the stationarity condition, the dependent variable would show signs of negative autocorrelation that often characterizes an over-differenced time series, whereas a value close to the upper boundary would induce stronger dependence due to a slower adjustment rate. In both cases, the strength of the cointegrating relationship diminishes and a natural question that arises is how the lasso-type methods handle such situations. Furthermore, when the adjustment rate is small in magnitude, e.g. $\alpha = -0.1$, the equilibrium correction may be so slow that for the purpose of forecasting it is best to model the data in differences regardless. In the following analysis we focus on the use of the adaptive lasso on a candidate set consisting of 50 variables and examine the effect of changes in the

Table 2.7 Cointegrated variables: the effect of α.

	Stationary			Non-stationary		
	$\alpha = -1.9$	$\alpha = -1.0$	$\alpha = -0.1$	$\alpha = -1.9$	$\alpha = -1.0$	$\alpha = -0.1$
Panel A: Levels						
RMSFE	**1.21**	**1.13**	1.09	**1.34**	1.25	0.38
MSFE	25.77	4.68	16.33	30.15	5.53	5.58
Consistent	31.7%	57.3%	14.5%	16.8%	7.9%	0.0%
Conservative	79.1%	97.0%	32.3%	59.8%	89.0%	12.8%
Variables	4.00	3.95	3.00	4.42	6.86	12.66
Panel B: ADF Differences						
RMSFE	3.54	2.14	0.14	3.48	1.73	0.14
MSFE	75.34	8.85	2.06	78.52	7.67	2.08
Consistent	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Conservative	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%
Variables	0.43	0.42	0.36	0.46	0.50	0.48
Panel C: Oracle Differences						
RMSFE	3.64	1.21	**0.08**	3.58	**1.17**	**0.08**
MSFE	77.48	5.03	1.16	80.74	5.18	1.23
Consistent	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Conservative	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Variables	1.95	0.57	0.38	0.41	0.38	0.43

Notes: see notes in Table 2.6. The evaluation metrics considered are listed in the first column. The models are estimated by the adaptive lasso with either (A) all variables in levels, (B) transformed variables based on the results of an ADF-test for stationarity or (C) infeasibly transformed variables based on knowledge of the true DGP.

adjustment rate on both the prediction and selection performance. For every adjustment rate, we examine the performance of the model estimated in three specifications; (1) all variables in the candidate set enter in levels, (2) some of the variables enter in differenced form based on the outcome of an Augmented Dickey-Fuller (ADF) test for stationarity of size 0.05, and (3) all variables that are simulated as I(1) variables enter the model in differenced form.[7] These models are listed in panel A, B and C of Table 2.7, respectively. The lowest RMSFE for a given adjustment rate across the three specification is denoted with bold font.

Models estimated in levels (panel A) only attain reasonable selection for an adjustment rate of $\alpha = -1$. Moving the adjustment rate towards the boundaries of the stationarity condition generally results in an increase in MSFE. However, different from the previous experiments, the strength of the adjustment rate also affects the OLS oracle estimator which serves as benchmark. A surprising finding is that the adaptive lasso does substantially better than the OLS oracle estimator when the adjustment rate is slow ($\alpha = -0.1$) and the candidate set contains irrelevant I(1)

[7] The effect of different strategies to pre-test for unit roots is examined in Chapter 5.

variables. We expect that the inclusion of a large number of unrelated random walks allows for a better in-sample fit resulting in a lower forecast error; since the reported forecasts are single step forecast, the improved in-sample fit may favour the predictive performance of the resulting spurious models, because the combined effect of the corresponding random coefficients is unlikely to push the prediction of the dependent variable far from its realized value. However, this statistical artefact cannot be expected to carry through to forecasts over longer horizons as the the trending behaviour of the I(1) variables will cause the predictions to drift away from the realisations. Indeed, in unreported analyses we find that the predictive superiority of the adaptive lasso on weakly cointegrated variables relative to the OLS oracle procedure vanishes at a forecast horizon of 10 steps and keeps deteriorating for longer horizons, as one would expect to be the case for forecasts with spurious regressions.

The models estimated on transformed data based on ADF-tests in panel B all obtain substantially higher RMSFEs, unless the equilibrium correction is small ($\alpha = -0.1$). Upon closer inspection, however, it becomes apparent that for these cases the adaptive lasso hardly incorporates any variables from the dataset, but rather forecasts the dependent variable by its time series average. The low RMSFEs obtained by this simple strategy imply that the use of cointegration with a slow adjustment rate has limited relevance for short-term forecasting purposes. Furthermore, for all adjustment rates the differenced models almost never contain all relevant variables. This provides an argument in favour of the use of L_1-penalized estimation in levels over the traditional approach of pre-processing the data, especially on datasets characterized by a "strong" cointegrating relationship ($\alpha = 1$). Finally, the infeasible models based on an oracle differencing procedure in panel C perform similar to the ADF-differenced data.

In conclusion, the use of lasso-type estimators on a high-dimensional non-stationary dataset containing cointegrated variables provides forecast gains over the traditional approach of using OLS on pre-processed data. A caveat to these results is that we rely on the underlying assumption of cointegration being present in the data. In practice, the uncertainty surrounding the validity of this assumption possibly affects the relative performance of the lasso-type methods. Accordingly, in the next chapter we propose a novel estimator that performs well regardless of whether cointegration is present.

2.4 Empirical Application

Complementing the simulation results, we perform an empirical application on a popular U.S. macroeconomic dataset. The dataset consists of 133 time series observed at a monthly frequency covering January 1959 to June 2015 and is obtained from the Fred-MD website.[8] In consideration of the potentially adverse consequences stemming from uncertainty regarding the presence of cointegration in empirical datasets, we postpone explicit modelling of cointegration to Chapter 5 and correct all series for non-stationarity. For the majority of series, this entails taking either log differences (e.g. real variables) or log second differences (e.g. price indices). Eight series are forecast, four of which are measures of real economic activity: real production income (RPI); total industrial production (IP); real manufacturing and trade sales (RMTS); and number of employees on non-agricultural payrolls (EMP). The remaining four series are price indices: the producer index for finished goods (PPI); the consumer price index (CPIA); the consumer price index less food (CPIUL); and the personal consumption expenditure implicit price deflator (PCEPI). These series, including their transformations, are similar to those frequently used in the seminal and contemporaneous forecasting literature (e.g. Stock and Watson, 2002b; Ludvigson and Ng, 2009; Kristensen, 2017).

The forecasts are generated as projections of an h-step-ahead variable y_{t+h}^h onto a set of variables observed up to time t that possibly includes lags of the dependent variable. As a benchmark, we consider a simple univariate AR model that obtains its forecasts by fitting the forecasting equation

$$y_{t+h}^h = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i+1} + \epsilon_{t+h}, \tag{2.16}$$

where y_{t+h}^h is defined appropriately according to the order of integration, see Stock and Watson (2002b) for details. The AR lag length p, for $p \in \{0, \dots, 6\}$, is determined by the BIC criterion, as is the case for all following methods. The penalized regressions obtain the forecasts by fitting

$$y_{t+h}^h = \alpha + \boldsymbol{x}_t' \boldsymbol{\beta}_x + \sum_{i=1}^{p} \beta_i y_{t-i+1} + \epsilon_{t+h}, \tag{2.17}$$

where the tuning parameters λ, α are selected using either the BIC, AIC or time series cross-validation. The autoregressive lags enter the model unpenalized across all specifications, their selection thus being dependent on the use of the BIC criterion rather

[8] https://research.stlouisfed.org/econ/mccracken/sel/

than the penalty induced shrinkage. Finally, forecasts based on static representations of factor models, i.e. all PC-type methods and the FHLR method, fit

$$y_{t+h}^h = \alpha + \hat{\boldsymbol{f}}_t' \boldsymbol{\beta}_f + \sum_{i=1}^{p} \beta_i y_{t-i+1} + \epsilon_{t+h}, \tag{2.18}$$

where the number of factors r is either kept fixed at five or determined by one of the information criteria of Bai and Ng (2002). Forecasts with the dynamic factor models FHLZ and DGR are based on

$$y_{t+h}^h = \alpha + \sum_{k=1}^{q} \hat{\boldsymbol{f}}_{t-k+1}^{*\prime} \boldsymbol{\beta}_{f,k} + \sum_{i=1}^{p} \beta_i y_{t-i+1} + \epsilon_{t+h}, \tag{2.19}$$

where $\hat{\boldsymbol{f}}_t^*$ is a s-dimensional vector of estimated dynamic factors. The number of lags of the factors that enter the forecast equation, $q \in \{0, \ldots, 6\}$, as well as the number of lags of the dependent variable are chosen by the BIC. We purposely do not forecast the target variable by iterated one-step ahead forecasts of the common and idiosyncratic components as is proposed in for example Forni et al. (2018), because the empirical performance of the iterated approach towards multi-step forecasts turned out to be highly inferior to the direct approach when forecasting the four price series. A similar finding is mentioned in Marcellino et al. (2006a) who consider the same series and compare direct and iterated forecasts with autoregressive models. While the detrimental effects of using iterated forecasts are slightly mitigated when modelling the price series as being I(1), the favourable performance for direct forecasts persists. Accordingly, we opt to model the price series as I(2) and report the results for the direct forecasts only.

We simulate real-time forecasting by calculating pseudo out-of-sample forecasts at horizons $h = 1$ and $h = 12$. An initial in-sample period covering 10 years of monthly observations is used to estimate the models by which to obtain the first out-of-sample prediction. For each new prediction, we keep the length of the in-sample period fixed and move the estimation sample forward by one period, i.e. we adopt a rolling window approach. The model is re-estimated prior to each prediction, including tuning parameter optimization, lag length selection, shrinkage and factor estimation. The forecasting performance is reported as the mean squared forecast error relative to the benchmark AR model. The comparison of forecasts is established based on the computation of Model Confidence Sets (MCS), as proposed by Hansen et al. (2011). We largely follow their original implementation with the $T_{R,\mathscr{M}}$-statistic and $\alpha = 0.25$. However, we do not adopt the moving-block bootstrap (MBB) procedure, given that

the time series of forecast errors display clear signs of unconditional heteroskedasticity over the full sample. Rather, we opt for the autoregressive wild bootstrap (AWB) which maintains its validity under the presence of both serial dependence and heteroskedasticity (Smeekes and Urbain, 2014a). The autoregressive coefficient (γ) that governs the amount of dependence captured in the AWB is determined by fitting individual MA models to the series of forecast errors with their individual order being chosen by the AIC criterion. We use the median order of the MA models (q) as a criterion for determining an appropriate block length, which we convert into the autoregressive coefficient with the conversion formula $\gamma = 0.01^{\frac{1}{q}}$ as proposed in Smeekes and Urbain (2014a, p.8). In a preliminary analyisis, however, we find that the use of the MBB generally results in model confidence sets that contain the same models as those generated with the AWB.

We visualize the Model Confidence Sets graphically for the 12-month ahead forecasts in Figure 2.2 while providing additional means of model comparisons with the use of the Diebold-Mariano tests in Figure 2.3. Comparisons of the monthly forecasts and a summary of the best performers are listed in the Appendix 2.A. The blue coloured bars in Figure 2.2 represent the models contained in the MCS, while the red bars are removed and are thus considered to be models with statistically inferior predictive capability for the respective series-horizon. In absolute terms, we observe that for the real series (left column) the factor models seem to outperform the lasso-type methods with PC, SPC, and FHLR showing strong performance in particular, whereas the lasso-type methods are comparable to the factor models for the nominal series (right column). The comparisons based on MCS almost always leaves all models in the set, seemingly suggesting that the variability in the forecast errors is too large to make any conclusive statements about the inferiority of certain models within the adopted 95% confidence level. The only exceptions to this are the exclusion of the lasso-type estimators for forecasts of Real Production Income (RPI) and occasionally some of the dynamic factor models FHLZ or DGR. An apparently counter-intuitive finding is that some of the methods removed from the MCS, e.g. the lasso in RPI, can have lower forecast losses than some of the models included in the MCS, e.g. "WPC-SWa" in RPI. The intuition behind this curiosity is that the series that, despite their higher MSFEs, are included in the MCS display higher variability in their forecast errors which prevents one from concluding that the method performs worse than other methods with certainty, although one may rightfully wonder whether it is desirable to consider models with higher average loss superior simply because they display larger variation in their loss. Additionally, by controlling the familywise error rate (FWE), that is, the probability of making a single false rejection, the power of the

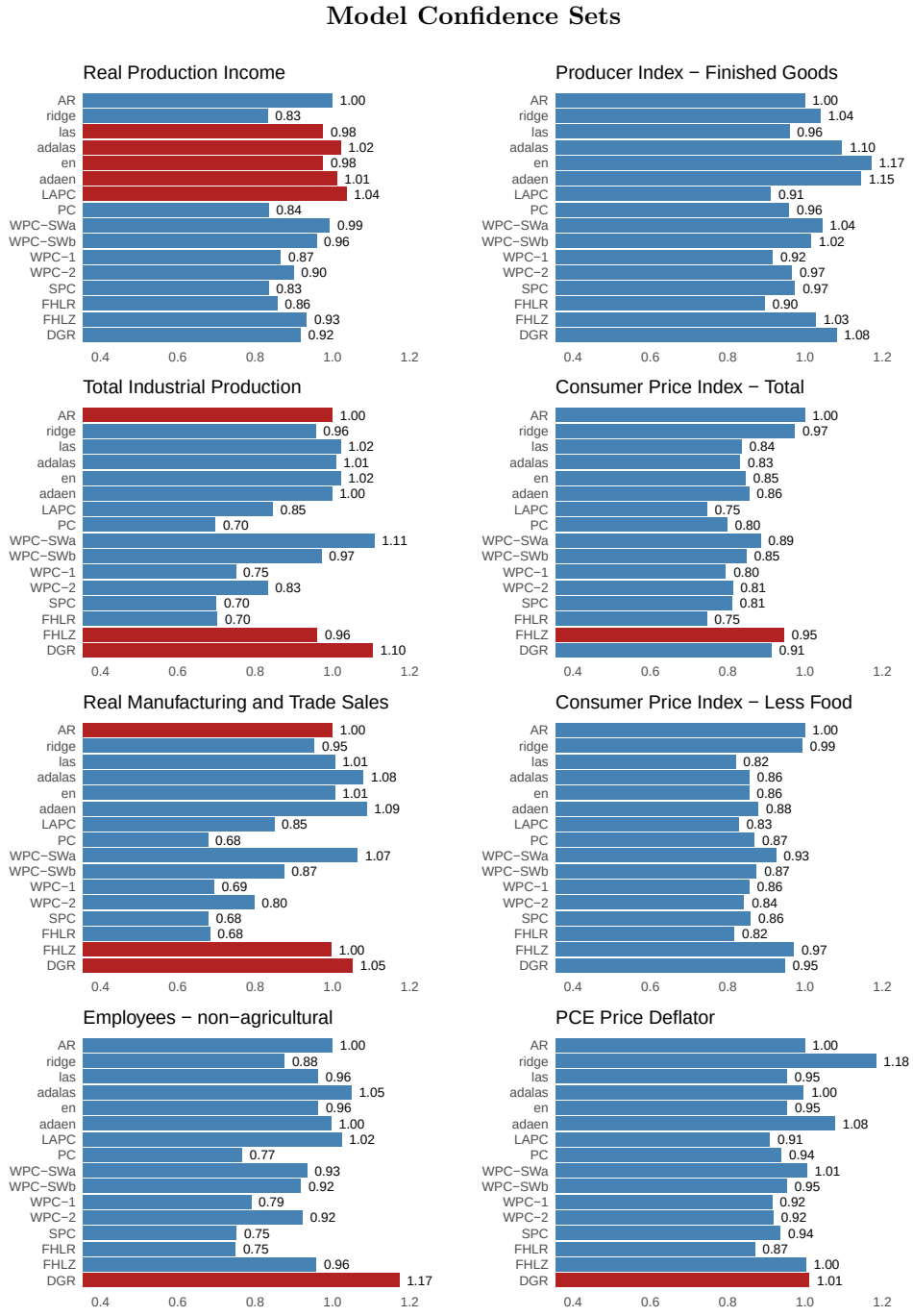

Figure 2.2: Blue coloured bars represent members of the Model Confidence Sets. Results are for 12-month ahead forecasts.

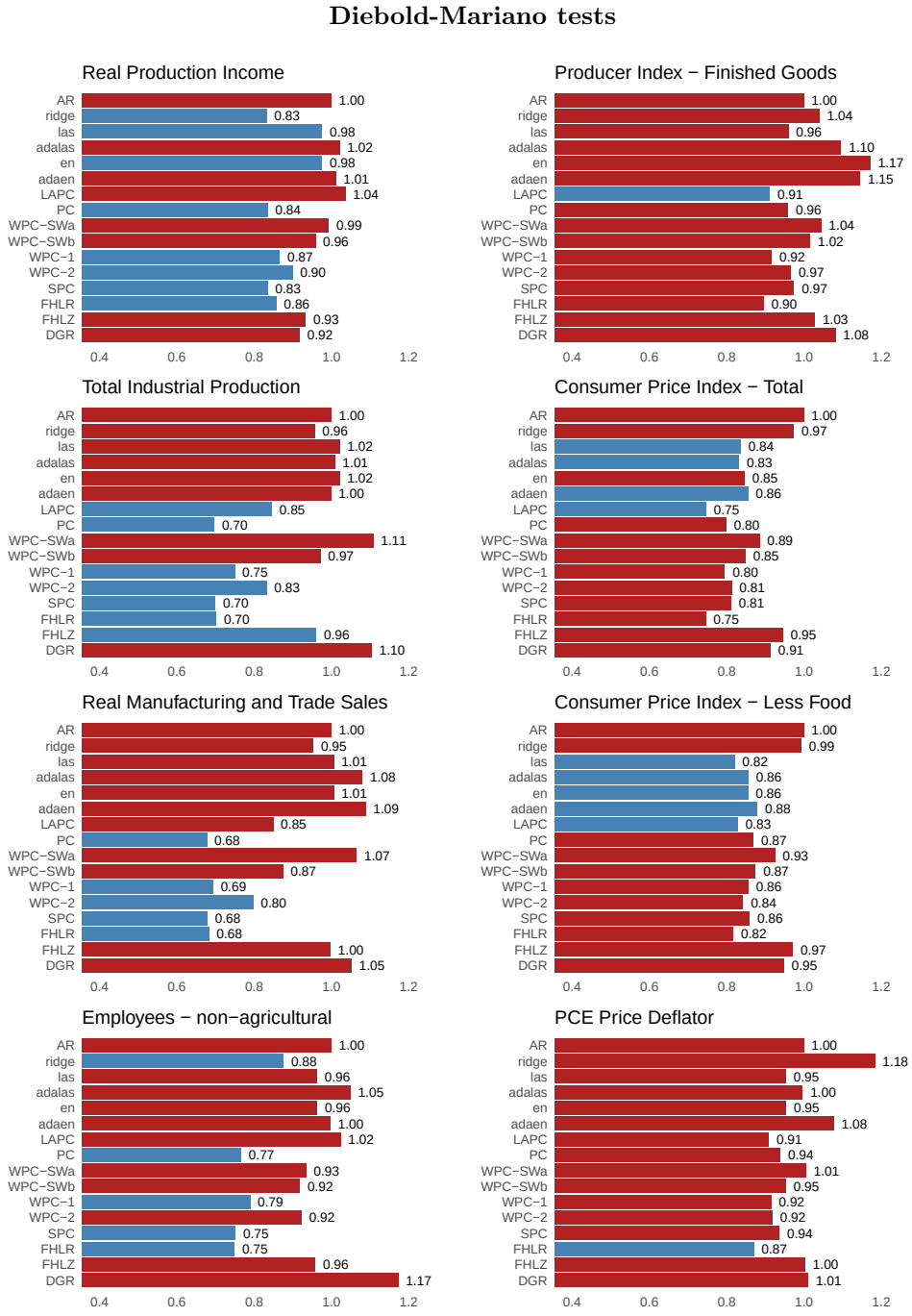

Figure 2.3: Blue coloured bars represent models with RMSFEs significantly less than 1. Results are for 12-month ahead forecasts.

MCS is highly dependent on the number of models considered. Our relatively large set of models is therefore detrimental in that respect. For this reason we also consider pairwise Diebold-Mariano tests which, by not controlling FWE, are not sensitive to this issue.

The Diebold-Mariano tests show frequent rejections of the null hypothesis of equal predictive capabilities in reference to the AR benchmark. The dominance of factor models on the real series and of the lasso-type estimators on the (consumer) price indices is immediately notable; on the real series most of the factor models are considered to obtain MSFEs significantly lower than the AR benchmark, whereas for the consumer price indices rejection only occurs for the methods involving L_1-shrinkage which is partially attributable to the lower variability in forecast errors of these methods. Finally, the dynamic factor methods FHLZ and DGR tend to perform slightly worse than the static variants, although we cautiously note that this may be a somewhat unfair comparison given the availability of a larger range of factor selection approaches for the static models. Indeed, during simulations we observed the Hallin and Liška criterion to occasionally deliver sub par performance. Given that the main comparison of interest, however, is the difference in predictive capability between shrinkage and factor methods we do not consider this caveat to impede our conclusions.

Hyperparameters and factor selection

We briefly comment on the performance of individual tuning methods for each model. The best performance by the shrinkage estimators is most frequently attained by tuning with the BIC criterion and CV coming in second place. For the static factor methods, the criteria most frequently leading to the best forecasting performance tend to be one of the Alessi et al. (2010) criteria, their IC3 criteria showing strong performance in particular. For the dynamic factor methods the use of a single dynamic factor performs best, followed by the use of four dynamic factors and the Hallin and Liška (2007) performs worst, possibly explaining the suboptimal predictive capability of the dynamic factor methods.[9] Lastly, the PC(LA) approach based on a preliminary lasso estimation performs similar when the lasso is tuned with either the BIC-criterion or the AIC-criterion.

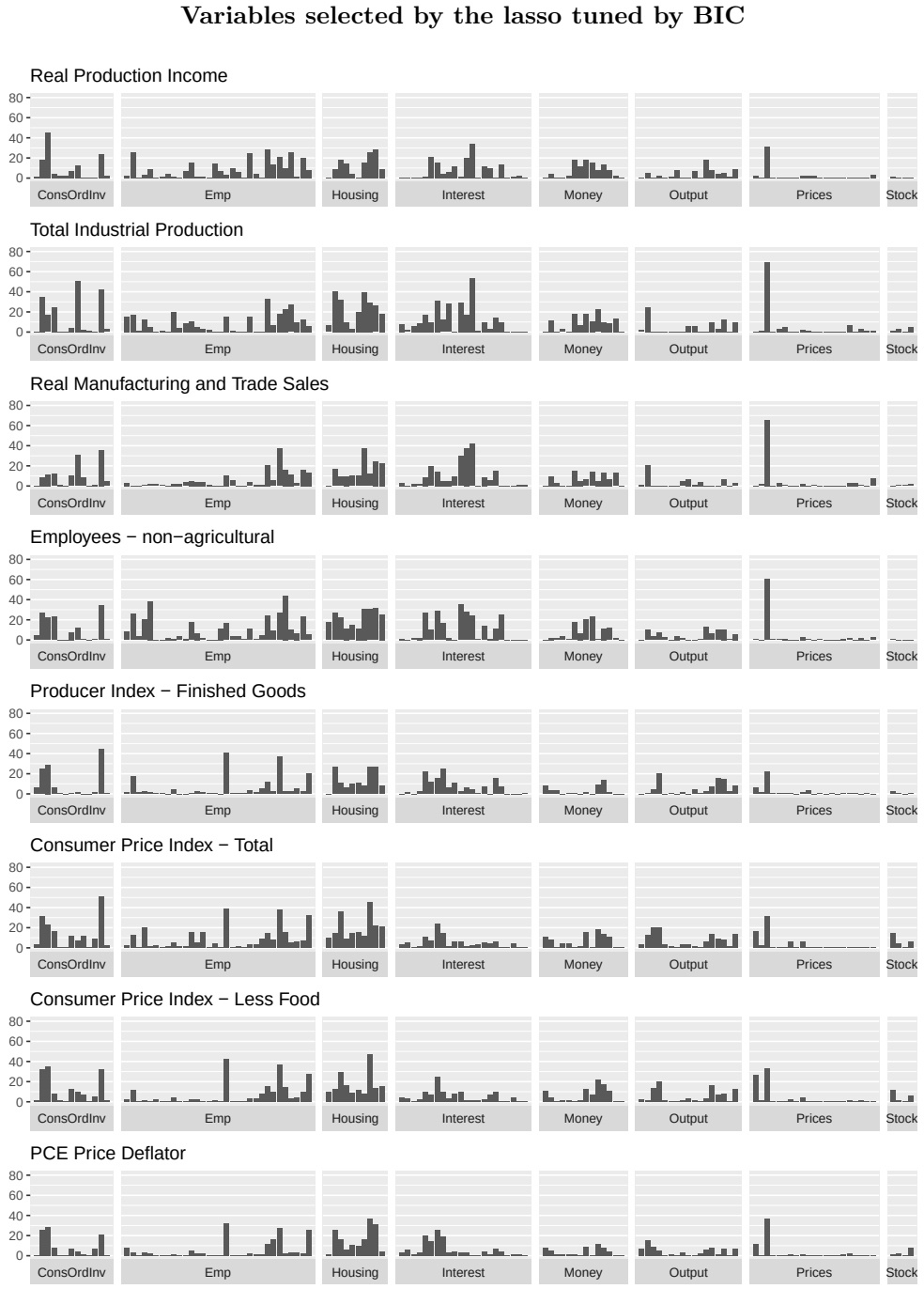

Figure 2.4: The percentage of times a variable is included in the forecast equation, separated by economic category.

Figure 2.5: An overview of the temporal selection properties per variable.

Variable selection and sparsity patterns

The documented performance of the lasso-type methods may leave one wondering whether the assumption of latent factors driving the variation in observable economic time series is justified. We explore the proposition of De Mol et al. (2008) where the collinearity induced by latent factors allows for approximation of the factor space with relatively few observable variable, while simultaneously resulting in highly unstable variable selection. In figure 2.4 we display the fraction of 12-month ahead forecast equations in which each variable in the data is selected by the lasso tuned with the BIC criterion. Strikingly, the pattern of frequently chosen variables is fairly consistent across the different forecast series, in particular when considering the group of nominal and real target variables separately. For example, in the Prices category, the "ISM Manufacturing: Price Index" (NAPMPRI) seems to capture the majority of the variation, whereas for the housing category the variables seem to substitute each other based on the low frequencies with which they are selected.[10] Not a single variable, however, is chosen consistently over all forecast periods. In line with the proposition of De Mol et al. (2008), this could be due to temporal instability resulting from collinearity induced by latent factors. Alternatively, structural changes may occur over the complete sample causing the relevance across variables to shift over time. To distinguish between these contrasting explanations we plot an overview of the variable selection over time in Figure 2.5, where a green bar indicates that the variable was included in the forecast while a red bar indicates exclusion. The vertical axis contains the 515 12-month ahead forecasts performed. Directly observable is the persistence in the selection of the most frequently included variables in the consumption, employment and prices categories, for which the structural change explanation seems most applicable. For other categories, such as housing or interest, factor-induced collinearity may offer an appropriate description, however.

The housing category provides a particularly suitable subset to examine whether the overlap in informational content of individual time series allows for approximation of the factor space with only a few cleverly selected variables. We focus on the 12-month ahead forecasts of Total Industrial Production (INDPRO) and consider the five most frequently chosen housing variables. We construct five new binary time series that indicate whether a variable for a given forecast at time $t+h$ was included and we refer to these as the selection series. Under the conjecture that the selection is unstable because the individual variables approximate the same space, one would

[9] We evaluate the Hallin and Liška criterion at three different sample points, i.e. (N_c, T_c) with $c \in \{1, 2, 3\}$, which is not necessarily optimal for the current empirical application.

[10] An overview of the most frequently chosen variable per economic category is provided in Table 2.8 in the Appendix.

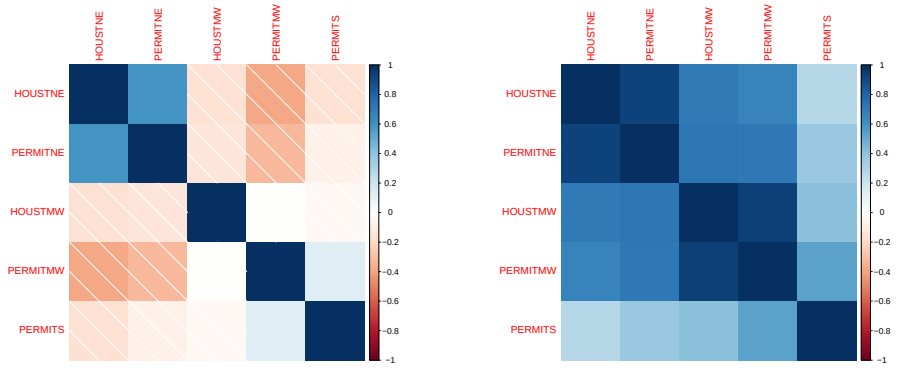

Figure 2.6: Plots of correlations in the selection series (left) and absolute correlations in the realizations (right) of the housing series most frequently selected in "INDPRO" forecasts.

expect to observe negative correlation between the selection series due to substitution effects and this negative correlation between the selection series should be stronger for time series that exhibit strong correlation in their realizations. Accordingly, we list two correlation plots in Figure 2.6. Evidence in favour of this conjecture would match up large negative correlation in the selection series, i.e. dark red boxes in the left plot, with large absolute correlations in the realizations of the respective series, i.e. dark blue boxes in the right plot. However, we observe that the selection series exhibit only mild negative correlation and the strongest correlated variables, i.e. "HOUSTNE" and "PERMITNE", actually tend to be selected together rather than substitute each other. We interpret these findings as anecdotal evidence that the variables selected by the lasso each contribute unique information and that structural change in the underlying DGP offers a feasible explanation of the temporal instability in the selection properties alongside the proposition of factor-induced collinearity in the observed time series.

2.5 Conclusion

In this chapter we examine the forecasting performance of (i) static, weighted and dynamic factor models, (ii) shrinkage estimators including ridge regression, the (adaptive) lasso and (adaptive) elastic-net and (iii) hybrid models in the form of a sparse principal components estimator and post-selection static factor models. Comprehensive simulations based on a wide variety of data generating processes indicate that lasso-type estimators are relatively robust against alternative DGP specifications; they naturally perform well on sparse and stationary models driven by ob-

served variables, but they also show strong forecasting performance on data driven by approximate factor structures, even when the latter models contain a high degree of non-sphericity in the idiosyncratic component. An empirical application on eight macroeconomic time series confirms the strong performance of factor-based models that is frequently covered in the forecasting literature. However, for certain target series such as the Consumer Price Index the lasso-type methods offer comparable if not better forecasting performance, while simultaneously displaying fairly persistent variable selection behaviour. We take this as further evidence that the assumption of common factors being persistent in macroeconomic data may not always be valid or, at a minimum, is not always relevant for forecasting purposes given the flexibility with which lasso-type estimators can handle this type of data.

A direct application of lasso-type estimators to a high-dimensional non-stationary dataset, in which the dependent variable is cointegrated with a small subset of the data, is shown to provide forecast improvements over the OLS estimator. However, we additionally find that a large number of irrelevant integrated variables are included when the model is specified in levels. Alternatively, when the data is transformed to stationarity by differencing, the estimators tend to exclude nearly all variable from the model. Hence, it is likely that the correct model specification lies somewhere in between these two extremes. This consists the topic of the next chapter.

Appendix 2.A One-Month Ahead Forecasts

Model Confidence Sets

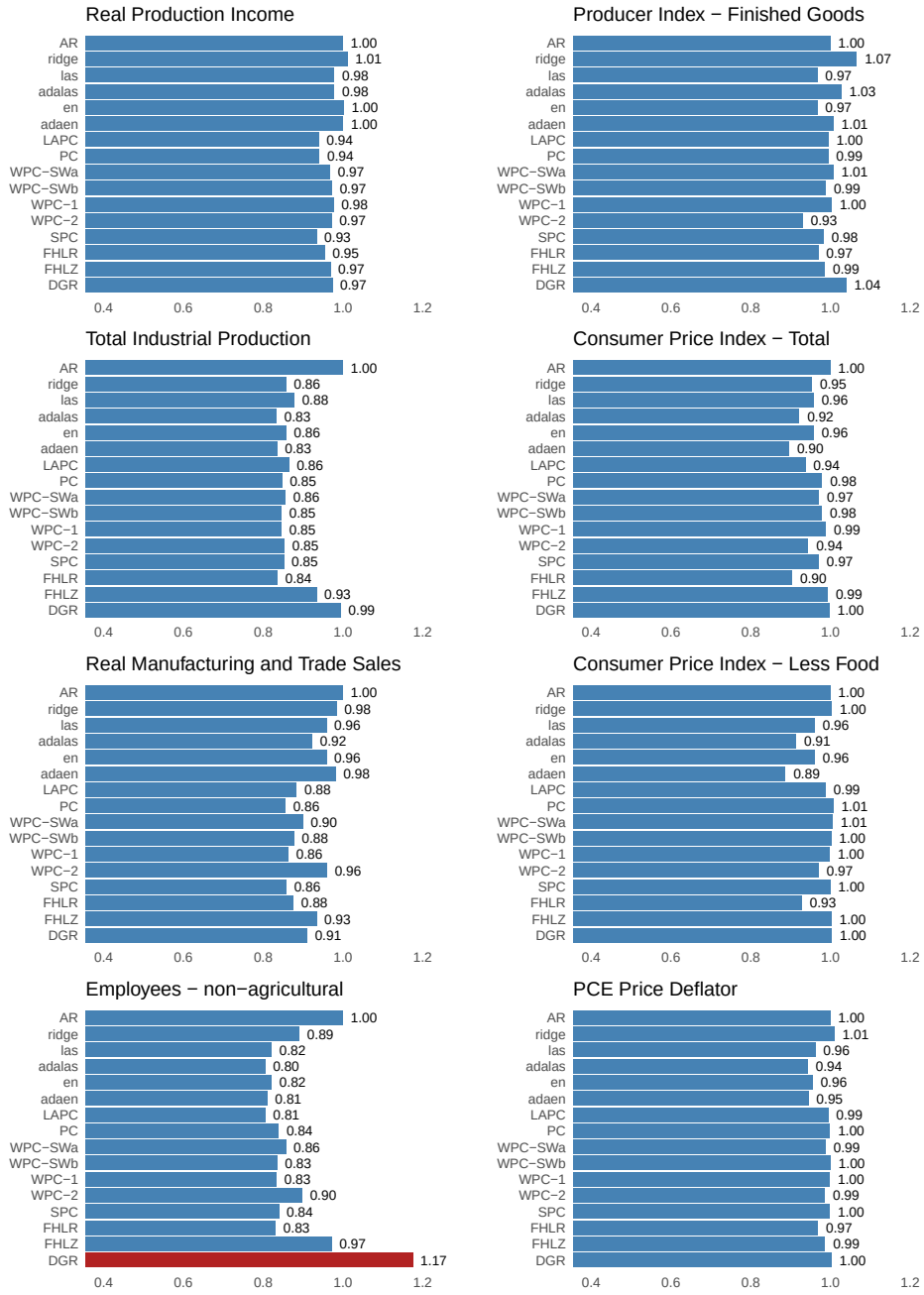

Figure 2.7: Blue coloured bars represent members of the Model Confidence Sets. Results are for 1-month ahead forecasts.

Diebold-Mariano Tests

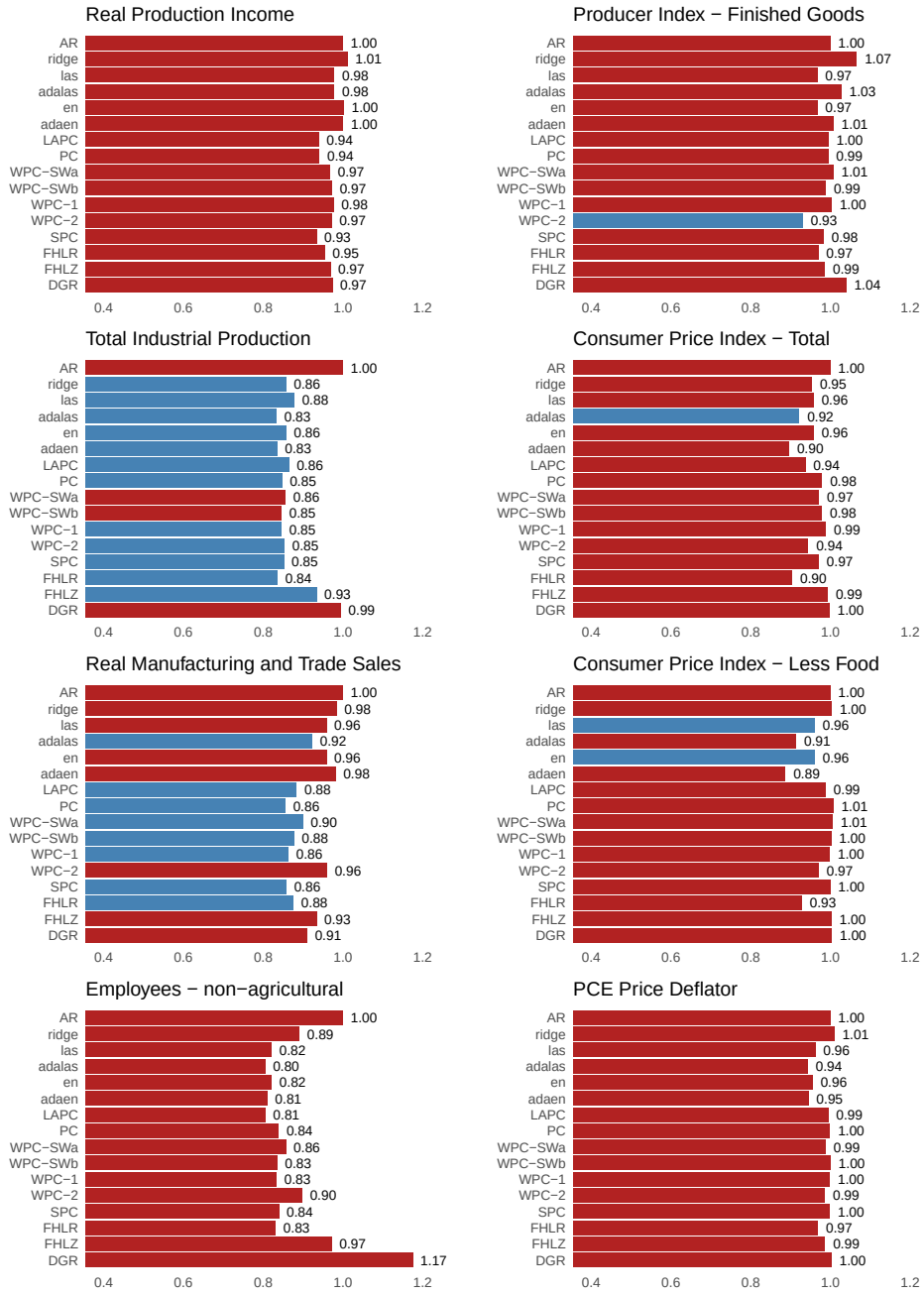

Figure 2.8: Blue coloured bars represent models with RMSFEs significantly less than 1. Results are for 1-month ahead forecasts.

Appendix 2.B Selected Variables

Table 2.8 Most Frequently Selected Variables

Forecast	ConsOrdInv	Emp	Housing	Interest
RPI	NAPMSDI	USWTRADE	PERMITS	BAAFFM
INDPRO	BUSINVx	USWTRADE	HOUSTNE	BAAFFM
CMRMTSPLx	M2REAL	USFIRE	PERMITNE	BAAFFM
PAYEMS	M2REAL	USGOVT	PERMITS	T10YFFM
PPIFGS	M2REAL	CES1021000001	PERMITS	TB6SMFFM
CPIAUCSL	M2REAL	CES1021000001	PERMITMW	TB3SMFFM
CPIULFSL	NAPMSDI	CES1021000001	PERMITMW	TB3SMFFM
PCEPI	NAPMSDI	CES1021000001	PERMITMW	TB3SMFFM
Forecast	Money	Output	Prices	Stock
RPI	CONSPI	IPBUSEQ	NAPMPRI	DTCOLNVHFNM
INDPRO	S.P.PE.ratio	W875RX1	NAPMPRI	INVEST
CMRMTSPLx	CONSPI	W875RX1	NAPMPRI	INVEST
PAYEMS	S.P.div.yield	IPBUSEQ	NAPMPRI	DTCOLNVHFNM
PPIFGS	FEDFUNDS	CMRMTSPLx	NAPMPRI	INVEST
CPIAUCSL	S.P.PE.ratio	DPCERA3M...	NAPMPRI	INVEST
CPIULFSL	S.P.PE.ratio	CMRMTSPLx	NAPMPRI	INVEST
PCEPI	S.P.PE.ratio	W875RX1	NAPMPRI	INVEST

Notes: this table report the most frequently selected variables in 12-month ahead forecast by the lasso tuned with the BIC criterion. For an overview of all the variables and their abbreviations, see the appendix in McCracken and Ng (2016)

Chapter 3

An Automated Approach Towards Sparse Single-Equation Cointegration Modelling

"Goodness is often defined in terms of prediction accuracy, but parsimony is another important criterion: simpler models are preferred for the sake of scientific insight into the x-y relationship."

- Efron, Hastie, Johnstone and Tibshirani (2004)

Abstract[†]

In this chapter we propose the Single-equation Penalized Error Correction Selector (SPECS) as an automated estimation procedure that directly incorporates the (co)integrating properties of the data. In Chapter 2, we documented favourable performance of penalized regression methods applied to stationary time series. However, by transforming the data to stationarity, we may lose predictive power and model interpretability by ignoring potential cointegration among the variables. Therefore, by extending the classical single-equation error correction model, SPECS enables the researcher to model large cointegrated datasets without necessitating any form of pretesting for the order of integration or cointegrating rank. We show that SPECS is able to consistently estimate an appropriate linear combination of the cointegrating vectors that may occur in the underlying DGP, while simultaneously enabling the correct recovery of sparsity patterns in the corresponding parameter space. A simulation study shows strong selective capabilities, as well as superior predictive performance in the context of nowcasting compared to high-dimensional models that ignore cointegration. An empirical application to nowcasting Dutch unemployment rates using Google Trends confirms the strong practical performance of our procedure.

[†]This chapter is based on Smeekes and Wijler (2018a).

3.1 Introduction

In this chapter we propose the Single-equation Penalized Error Correction Selector (SPECS) as a tool to perform automated modelling of a potentially large number of time series of unknown order of integration. In many economic applications, datasets will contain possibly (co)integrated time series, which has to be taken into account in the statistical analysis. Traditional approaches include modelling the full system of time series as a vector error correction model (VECM), estimated by methods such as maximum likelihood estimation (Johansen, 1995a), or transforming all variables to stationarity before performing further analysis. However, both methods have considerable drawback when the dimension of the dataset increases.

While the VECM approach allows for a general and flexible modelling of potentially cointegrated series, and the optimality properties of a correctly specified full-system estimator are theoretically attractive, these estimators suffer from the curse of dimensionality due to the large number of parameters to estimate. In practice they therefore quickly become difficult to interpret and computationally intractable on even moderately sized datasets. As such, to reliably apply such full-system estimators requires non-trivial a priori choices on the relevance of specific variables to keep the dimension manageable. Moreover, in many cases of practical relevance, one only has a single variable of interest, and estimating the parameter-heavy full system is not necessary. On the other hand, the alternative strategy of prior transformations to stationarity is more easily compatible with single variables of interest and larger dimensions, but requires either a priori knowledge of the order of integration of individual variables, or pre-testing for unit roots, which is prone to errors in particular if the number of variables is large. Additionally, this approach ignores the presence of cointegration among the variables, which may have detrimental effects on the subsequent analysis. In an attempt to resolve these issues, we propose SPECS as an alternative approach towards intuitive automated modelling of large non-stationary datasets.

SPECS is a form of penalized regression designed to sparsely estimate a conditional error correction model (CECM). We demonstrate that SPECS possesses the oracle property as defined in Fan and Li (2001) in a fixed-dimensional asymptotic framework.[1] In particular, SPECS simultaneously allows for consistent estimation of

[1] The choice for a fixed-dimensional framework is based on expositional simplicity. This framework allows us to introduce our estimator under a set of intuitive assumptions and to elaborately discuss additional issues such as weak exogeneity and mixed orders of integration. In Chapter 4 we extend the results to a high-dimensional framework.

the non-zero coefficients and the correct recovery of sparsity patterns in the single-equation model. It therefore provides a fully data-driven way of selecting the relevant variables from a potentially large dataset of (co)integrated time series. Moreover, due to the flexible specification of the single-equation model, SPECS is able to take into account cointegration in the dataset without requiring any form of pre-testing for unit roots or testing for the cointegrating rank, and can thus be applied "as is" to any dataset containing an (unknown) mix of stationary and integrated time series. As a companion to this chapter, ready-to-use *R* code is available online that implements an intuitive and easy-to-interpret algorithm for SPECS estimation.[2]

Single-equation error correction models are frequently employed in tests for cointegration (e.g. Engle and Granger, 1987; Phillips and Ouliaris, 1990; Boswijk, 1994; Banerjee et al., 1998) as well as in forecasting applications (e.g. Engle and Yoo, 1987; Chou et al., 1996), but require a weak exogeneity assumption for asymptotically efficient inference (Johansen, 1992a). Weak exogeneity entails the existence of a single cointegrating vector that only appears in the marginal equation for the variable of interest. If this assumption holds, our procedure can be interpreted as an alternative to cointegration testing in the ECM framework (Boswijk, 1994; Palm et al., 2010). However, weak exogeneity may not be realistic in large datasets and we provide detailed illustrations of the implications of failure of this assumption and demonstrate that absent of weak exogeneity our procedure consistently estimates a linear combination of the true cointegrating vectors. While this impedes inference on the cointegrating relations, when the main aim of the model is nowcasting or forecasting, our procedure remains theoretically justifiable and provides empirical researchers with a simple and powerful tool for automated analysis of high-dimensional non-stationary datasets. In addition, for modeling a single variable of interest using a large set of potential regressors, SPECS provides a variable selection mechanism, allowing the researcher to discard variables that are irrelevant for this particular analysis. Our simulation results demonstrate strong selective capabilities in both low and high dimensions. Furthermore, a simulated nowcasting application highlights the importance of incorporating cointegration in the data as our proposed estimators obtain higher nowcast accuracies in comparison to a penalized autoregressive distributed lag (ADL) model. This finding is confirmed in an empirical application, where SPECS is employed to nowcast Dutch unemployment rates with the use of a dataset containing Google Trends series.

Recent literature has also seen the development of methods for analyzing high-dimensional (co)integrated time series. Kock (2016) proposes the adaptive lasso to

[2] https://sites.google.com/view/etiennewijler

estimate an augmented Dickey-Fuller regression. While this univariate model is inherently different from ours, it provides an insightful demonstration of how the lasso may be used as an alternative to testing for non-stationarity, paralleling our suggestion to consider SPECS as an alternative for cointegration testing under the assumption of weak exogeneity.

For VECM systems, Wilms and Croux (2016) propose a penalized maximum likelihood approach, with shrinkage performed on the cointegrating vectors, the coefficients regulating the short-run dynamics and the covariance matrix. While their method is shown to obtain forecast gains relative to the traditional Johansen method, no theoretical results are provided. Liao and Phillips (2015) provide an automated method of joint rank selection and parameter estimation with the use of an adaptive penalty and derive oracle properties in a fixed-dimensional framework. Next to this theoretical limitation on its applicability to large datasets, practical implementation is further complicated due to reliance on the eigenvalue decomposition of an asymmetric matrix, which introduces complex values into the corresponding objective function. As noted by Liang and Schienle (2019, p. 424), this results in a non-standard harmonic function optimization problem. Liang and Schienle (2019) propose joint parameter estimation and rank determination by employing a penalty that makes use of the QR-decomposition of the long-run coefficient matrix. This method possesses oracle-like properties under a high-dimensional asymptotic regime, but it requires the availability of an initial OLS estimator, thereby preventing applications on datasets in which the number of variables exceeds, or is close to, the number of available time series observations. Additionally, estimation of the long-run and short-run dynamics is performed sequentially rather than simultaneously, necessitating a two-step procedure.

In a single-equation setting, Lee et al. (2018) derive fixed-dimensional oracle properties for the adaptive lasso applied to predictive regressions where the regressors are allowed to be of mixed orders of integration. However, as a consequence of their model formulation in which all variables enter in levels, their estimator appears to be susceptible to spurious regression when the regressors are not cointegrated.

Finally, outside the penalized regression framework, Zhang et al. (2019a) propose an eigenvalue decomposition to estimate the cointegrating space in the presence of any integer and fractional order of integration of the variables. However, the estimation procedure proposed by Zhang et al. does not perform variable selection, nor does it provide explicit estimates of the transient dynamics in a VECM. Onatski and Wang (2019) develop a novel inference procedure for the cointegrating rank in high dimensions. Similar to the Johansen procedure, their test is based on the squared

canonical correlations, for which they derive the limit spectral distribution under joint asymptotics with the use of arguments from random matrix theory.

Our proposed method provides several contributions to this existing literature. First, unlike many of the penalized regression methods surveyed above, the practical implementation of SPECS is straightforward for large datasets, including cases where the number of parameters is larger than the time dimension. Second, our method completely removes the need for pre-testing for the order of integration or cointegrating rank, and is not sensitive to spurious regression. Third, to the best of our knowledge, our paper is the first to explicitly allow for the presence of deterministic components in the theory, a crucial feature for many applications. Fourth, in the next chapter we extend our theoretical results to a high-dimensional framework where the number of parameters is allowed to grow with the sample size. This requires non-standard theoretical results on bounds of the smallest eigenvalue of a matrix of (co)integrated regressors, similar to those in Zhang et al. (2019a), which are further developed in Chapter 4.

The chapter is structured as follows. In Section 3.2 we discuss the data generating process and describe the SPECS estimator. The main theoretical results of the chapter are presented in Section 3.3. Section 3.4 contains several simulation studies, followed by an empirical application in Section 3.5. We conclude in Section 3.6. Proofs of the main results are presented in Appendix 3.A and additional results are contained in Appendix 3.B.

Notation

Finally, a word on notation. We use $\|\cdot\|_p$ to denote the L_p-norm, i.e. $\|\boldsymbol{v}\|_p = (\sum_{i=1}^n |v|_i^p)^{1/p}$ for a vector $\boldsymbol{v} \in \mathbb{R}^n$ and $\|\boldsymbol{V}\|_p = \left(\sum_{j=1}^m \sum_{i=1}^n |v|_{ij}^p\right)^{1/p}$ for a matrix $\boldsymbol{V} \in \mathbb{R}^{n \times m}$. The maximum (minimum) elements of a matrix $\boldsymbol{A}$ is denoted by $a_{\max}$ ($a_{\min}$), and we use $\boldsymbol{A} \succ 0$ to denote that the matrix is positive definite. In addition, we let $\boldsymbol{A}_\perp$ denote the orthogonal complement of $\boldsymbol{A}$, such that $\boldsymbol{A}_\perp' \boldsymbol{A} = 0$. If $\boldsymbol{v}$ is a sparse vector and $\boldsymbol{u}$ is another vector of similar dimension, we define the support index of $\boldsymbol{v}$ as $S_v = \{i | v_i \neq 0\}$ and $\boldsymbol{u}_{S_v}$ as the sub-vector of $\boldsymbol{u}$ indexed by S_v. Similarly, for a matrix $\boldsymbol{A}$, we use $\boldsymbol{A}_{S_v}$ to denote the matrix derived from $\boldsymbol{A}$, containing the columns indexed by S_v. We use a similar notation for the complement of the support, i.e S_v^c, $\boldsymbol{u}_{S_v^c}$ and $\boldsymbol{A}_{S_v^c}$. Finally, convergence in distribution (probability) is denoted by $\xrightarrow{d}$ ($\xrightarrow{p}$).

3.2 The Single-Equation Penalized Error Correction Selector

3.2.1 Setup

Throughout the chapter we let our single variable of interest be denoted by y_t, which we aim to model dynamically with the use of an N-dimensional time series $\boldsymbol{z}_t = (y_t, \boldsymbol{x}_t')'$, described by

$$\boldsymbol{z}_t = \boldsymbol{\mu} + \boldsymbol{\tau} t + \boldsymbol{\zeta}_t, \tag{3.1}$$

with the stochastic component given by

$$\Delta\boldsymbol{\zeta}_t = \boldsymbol{A}\boldsymbol{B}'\boldsymbol{\zeta}_{t-1} + \sum_{j=1}^{p} \boldsymbol{\Phi}_j \Delta\boldsymbol{\zeta}_{t-j} + \boldsymbol{\epsilon}_t, \tag{3.2}$$

where $\boldsymbol{\epsilon}_t = (\epsilon_{1,t}, \boldsymbol{\epsilon}_{2,t}')'$. The model can be rewritten into a VECM form by substituting (3.1) into (3.2) to obtain

$$\Delta\boldsymbol{z}_t = \boldsymbol{A}\boldsymbol{B}' \left(\boldsymbol{z}_{t-1} - \boldsymbol{\mu} - \boldsymbol{\tau}(t-1)\right) + \tau^* + \sum_{j=1}^{p} \boldsymbol{\Phi}_j \Delta\boldsymbol{z}_{t-j} + \boldsymbol{\epsilon}_t, \tag{3.3}$$

where $\tau^* = (I - \sum_{j=1}^{p} \boldsymbol{\Phi}_j)\boldsymbol{\tau}$. From this representation, it can directly be observed that the presence of a constant in (3.1) results in a constant within the cointegrating relationship if $\boldsymbol{B}'\boldsymbol{\mu} \neq 0$. Furthermore, the linear trend in (3.1) appears as a constant in the differenced series and may additionally appear as a trend within the cointegrating vector if $\boldsymbol{B}'\boldsymbol{\tau} \neq \boldsymbol{0}$, the latter implying that the equilibrium error $\boldsymbol{B}'\boldsymbol{z}_t$ is a trend stationary process.

We impose the following assumption on the innovations.

Assumption 3.1. $\{\boldsymbol{\epsilon}_t\}_{t\geq 1}$ is an N-dimensional martingale difference sequence with $\mathbb{E}\left(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t'\right) = \boldsymbol{\Sigma} \succ 0$ and $\mathbb{E}\left|\boldsymbol{\epsilon}_t\right|^{2+\eta} < \infty$ for $\eta > 0$.

Under this assumption, the innovations satisfy the multivariate invariance principle

$$T^{-1/2} \sum_{t=1}^{[T\cdot]} \boldsymbol{\epsilon}_t \to \boldsymbol{B}(\cdot), \tag{3.4}$$

where $\boldsymbol{B}(\cdot)$ represents a vector Brownian motion with covariance matrix $\boldsymbol{\Sigma}$ (Phillips and Solo, 1992, p. 983).

For the VECM model to admit a vector moving average (VMA) representation we maintain the following assumptions.

Assumption 3.2. Define $\boldsymbol{A}(z) := (1-z) - \boldsymbol{A}\boldsymbol{B}'z - \sum_{j=1}^{p} \boldsymbol{\Phi}_j(1-z)z^j$.

(i) The determinantal equation $|\boldsymbol{A}(z)|$ has all roots on or outside the unit circle.

(ii) $\boldsymbol{A}$ and $\boldsymbol{B}$ are $N \times r$ matrices with $r \leq N$ and $\operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{B}) = r$. For $r = 0$, we adopt the convention that $\boldsymbol{A}\boldsymbol{B}' = 0$ and $\boldsymbol{A}_\perp = \boldsymbol{B}_\perp = \boldsymbol{I}_N$.

(iii) The $((N-r) \times (N-r))$ matrix $\boldsymbol{A}'_\perp \left(\boldsymbol{I}_N - \sum_{j=1}^{p} \boldsymbol{\Phi}_j\right) \boldsymbol{B}_\perp$ is invertible.

The importance in deriving a single-equation model for y_t, our main variable of interest, is to ensure that the variables modelling the variation in y_t remain exogenous. This is accomplished by orthogonalizing the errors driving the single-equation model, say $\epsilon_{y,t}$, from the errors driving the marginal equation of the endogenous variables x_t. Orthogonalization is achieved by decomposing $\epsilon_{1,t}$ into its best linear prediction based on $\boldsymbol{\epsilon}_{2,t}$ and the corresponding orthogonal prediction error. To this end, partition the covariance matrix of $\boldsymbol{\epsilon}_t$ as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \boldsymbol{\sigma}_{12} \\ \boldsymbol{\sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \tag{3.5}$$

such that we obtain

$$\epsilon_{1,t} = (0, \pi_0')\boldsymbol{\epsilon}_t + (1, -\boldsymbol{\pi}_0')\,\boldsymbol{\epsilon}_t = \hat{\epsilon}_{1,t} + \epsilon_{y,t} \tag{3.6}$$

where $\hat{\epsilon}_{1,t} = \boldsymbol{\pi}_0'\boldsymbol{\epsilon}_{2,t}$ with $\boldsymbol{\pi}_0 = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}$ and $\epsilon_{y,t} = (1, -\boldsymbol{\pi}_0')\,\boldsymbol{\epsilon}_t$ with $\mathbb{E}(\boldsymbol{\epsilon}_{2,t}\epsilon_{y,t}) = \boldsymbol{0}$ by construction. Writing out (3.6) in terms of the observable time series results in the single-equation model

$$\begin{aligned} \Delta y_t &= (1, -\boldsymbol{\pi}_0') \left(\boldsymbol{A}\boldsymbol{B}' \left(\boldsymbol{z}_{t-1} - \boldsymbol{\mu} - \boldsymbol{\tau}(t-1)\right) + \boldsymbol{\tau}^* + \sum_{j=1}^{p} \boldsymbol{\Phi}_j \Delta \boldsymbol{z}_{t-j} \right) \\ &\quad + \boldsymbol{\pi}_0' \Delta \boldsymbol{x}_t + \epsilon_{y,t} \\ &= \boldsymbol{\delta}' \boldsymbol{z}_{t-1} + \boldsymbol{\pi}' \boldsymbol{w}_t + \mu_0 + \tau_0 (t-1) + \epsilon_{y,t}, \end{aligned} \tag{3.7}$$

where $\boldsymbol{\delta}' = (1, -\boldsymbol{\pi}_0')\,\boldsymbol{A}\boldsymbol{B}'$, $\boldsymbol{\pi} = (\boldsymbol{\pi}_0', \ldots, \boldsymbol{\pi}_p')'$ with $\boldsymbol{\pi}_j = (1, -\boldsymbol{\pi}_0')\boldsymbol{\Phi}_j$ for $j = 1, \ldots, p$, $\boldsymbol{w}_t = (\Delta \boldsymbol{x}_t', \Delta \boldsymbol{z}_{t-1}', \ldots, \Delta \boldsymbol{z}_{t-p}')'$, $\mu_0 = (1, -\boldsymbol{\pi}_0')(\boldsymbol{\tau}^* - \boldsymbol{A}\boldsymbol{B}'\boldsymbol{\mu})$, $\tau_0 = -(1, -\boldsymbol{\pi}_0')\boldsymbol{A}\boldsymbol{B}'\boldsymbol{\tau}$, and $\epsilon_{y,t} = (1, -\boldsymbol{\pi}_0')\boldsymbol{\epsilon}_t$.

Remark 3.1. The single-equation model may alternatively be derived under the assumption of normally distributed errors. In this framework, $\epsilon_{y,t}$ has the conditional

normal distribution from which (3.7) can be obtained (cf. Boswijk, 1994). A benefit of assuming normality is that, under the additional assumption of weak exogeneity, the OLS estimates of (3.7) are optimal in the mean-squared sense. However, the assumption of normality is unnecessarily restrictive when the, perhaps overly, ambitious goal of complete and correct specification is abandoned.

In general, the implied cointegrating vector $\boldsymbol{\delta}$ in the single-equation model for y_t contains a linear combination of the cointegrating vectors in $\boldsymbol{B}$ with their weights being given by $(1, -\boldsymbol{\pi}_0')\,\boldsymbol{A}$. Since the marginal equations of $\boldsymbol{x}_t$ contain information about the cointegrating relationship, efficient estimation within the single-equation model is only attained under an assumption of weak exogeneity. Johansen (1992a) shows that sufficient conditions for weak exogeneity to hold are (i) $\boldsymbol{\epsilon}_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, (ii) $\operatorname{rank}(\boldsymbol{AB}') = 1$, i.e. there is a single cointegrating N-dimensional cointegrating vector $\boldsymbol{\beta}$, and (iii) the vector of adjustment rates takes on the form $\boldsymbol{\alpha} = (\alpha_1, \mathbf{0}')'$. However, these conditions are rather restrictive when considering high-dimensional economic datasets that are likely to possess multiple cointegrating relationships and complex covariance structures across the errors. Accordingly, we opt to derive our results without assuming weak exogeneity, while acknowledging that direct interpretation of the estimated cointegrating vector will only be valid in the presence of weak exogeneity.

3.2.2 Estimation Procedure

We propose to estimate (3.7) with penalized regression based on an L_1-penalty to attain sparse solutions. However, a property of L_1-penalized regression is that its solutions are not equivariant to arbitrary scaling of the variables, which is why the convention is to standardize the data prior to estimation (see Hastie et al., 2008, p. 8). While this practice is fairly innocuous in the stationary setting, this is not the case when dealing with non-stationary variables, as the standard variance estimates are diverging such that care has to be taken when deriving the asymptotic theory. Let $\boldsymbol{Z}_{-1} = (\boldsymbol{z}_0, \ldots, \boldsymbol{z}_{T-1})'$, $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T)'$, and write $\boldsymbol{V} = (\boldsymbol{Z}_{-1}, \boldsymbol{W})$, $\boldsymbol{\gamma} = (\boldsymbol{\delta}', \boldsymbol{\pi}')'$, $\boldsymbol{\theta} = (\mu_0, \tau_0)'$ and $\boldsymbol{D} = (\boldsymbol{\iota}, \bar{\boldsymbol{t}})$, where $\boldsymbol{\iota}$ is an N-dimensional vector of ones and $\bar{\boldsymbol{t}} = (0, \ldots, T-1)'$. For any data matrix $\boldsymbol{A}$, coefficient vector $\boldsymbol{b}$ and diagonal weighting matrix $\boldsymbol{\Sigma}_A$, define $\tilde{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{\Sigma}_A^{-1}$ and $\boldsymbol{b}^s = \boldsymbol{\Sigma}_A b$. Then, we can rewrite (3.7) in standardized matrix form as

$$
\begin{aligned}
\Delta \boldsymbol{y} &= \boldsymbol{Z}_{-1}\boldsymbol{\delta} + \boldsymbol{W}\boldsymbol{\pi} + \boldsymbol{\iota}\mu_0 + \bar{\boldsymbol{t}}\tau_0 + \epsilon_y \\
&= \boldsymbol{Z}_{-1}\boldsymbol{\Sigma}_Z^{-1}\boldsymbol{\Sigma}_Z\boldsymbol{\delta} + \boldsymbol{W}\boldsymbol{\Sigma}_W^{-1}\boldsymbol{\Sigma}_W\boldsymbol{\pi} + \boldsymbol{\iota}\mu_0 + \bar{\boldsymbol{t}}\tau_0 + \epsilon_y \\
&= \tilde{\boldsymbol{Z}}_{-1}\boldsymbol{\delta}^s + \tilde{\boldsymbol{W}}\boldsymbol{\pi}^s + \boldsymbol{\iota}\mu_0 + \bar{\boldsymbol{t}}\tau_0 + \epsilon_y = \tilde{\boldsymbol{V}}\boldsymbol{\gamma}^s + \boldsymbol{D}\boldsymbol{\theta} + \epsilon_y.
\end{aligned} \tag{3.8}
$$

We then estimate (3.8) with our shrinkage estimator, by minimizing the objective function

$$G_T\left(\boldsymbol{\gamma}^s, \boldsymbol{\theta}\right) = \left\| \Delta \boldsymbol{y} - \tilde{\boldsymbol{V}} \boldsymbol{\gamma}^s - \boldsymbol{D}\boldsymbol{\theta} \right\|_2^2 + P_\lambda(\boldsymbol{\gamma}^s). \tag{3.9}$$

The penalty function in (3.9) takes on the form

$$P_\lambda(\boldsymbol{\gamma}^s) = \lambda_{G,T} \left\| \boldsymbol{\delta}^s \right\|_2 + \lambda_{\delta,T} \sum_{i=1}^{N} \omega_{\delta,i}^{k_\delta} \left| \delta_i^s \right| + \lambda_{\pi,T} \sum_{j=1}^{M} \omega_{\pi,j}^{k_\pi} \left| \pi_j^s \right|, \tag{3.10}$$

where $\omega_{\delta,i}^{k_\delta} = 1/\left|\hat{\delta}_{Init,i}\right|^{k_\delta}$ and $\omega_{\pi,j}^{k_\pi} = 1/\left|\hat{\pi}_{Init,j}\right|^{k_\pi}$. The tuning parameters k_δ and k_π regulate the degree to which the initial estimates affect the penalty weights, and they should satisfy certain constraints that are specified in the theorems to follow. Throughout this chapter we assume that the initial estimators are $\sqrt{T}$-consistent; for example we can use $\hat{\delta}_{OLS}$ and $\hat{\pi}_{OLS}$.[3]

We denote the minimizers of (3.9) by $\hat{\boldsymbol{\gamma}}^s$ and $\hat{\boldsymbol{\theta}}$ and the de-standardized minimizers by $\hat{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}_V^{-1} \hat{\boldsymbol{\gamma}}^s$. The group penalty, regulated by $\lambda_{G,T}$, serves to promote exclusion of the lagged levels as a group when there is no cointegration present in the data. In this case, the model is effectively estimated in differences and corresponds to a conditional model derived from a vector autoregressive model specified in differences. The individiual L_1-penalties, regulated by $\lambda_{\delta,T}$ and $\lambda_{\pi,T}$ serve to enforce sparsity in the coefficient vector $\boldsymbol{\delta}$ and $\boldsymbol{\pi}$ respectively. Furthermore, the penalties are weighted by an initial estimator to enable simultaneous estimation and selection consistency of the coefficients. Note that the deterministic components μ_0 and τ_0 are left unpenalized, as their inclusion in the model is desirable to enable identification of the limiting distribution of the estimators. As shown in Yamada (2017), the inclusion of an unpenalized constant and deterministic trend is equivalent to de-meaning and de-trending the data prior to estimation.

Remark 3.2. SPECS incorporates an L_2 penalty to achieve sparsity on $\boldsymbol{\delta}$ at the group level, while inclusion of L_1 penalties ensures sparsity within and outside the group. The resulting optimization problem resembles that of the Sparse-Group Lasso (Simon et al., 2013), and the same algorithm can be employed here with only minor adjustments that account for the presence of just a single group. The R code that we make available online implements this algorithm to compute SPECS.

Remark 3.3. Standardization of unpenalized components does not affect the esti-

[3] In principal any consistent estimator would suffice, although the required growth rates of the penalty parameters in (3.10) are intrinsically related to the rate of convergence of the initial estimator.

mation of penalized components; a feature that can be directly verified by application of Lemma 3.A.4 in Appendix 3.A.1. Accordingly, we do not explicitly standardize the subset $\boldsymbol{D}$ containing the (deterministic) variables that are left unpenalized.

3.3 Theoretical Properties

In this section we derive the theoretical properties of SPECS. First, we establish the consistency and oracle properties of SPECS in Section 3.3.1. Thereafter, we consider the implications for particular model specifications in Section 3.3.2.

3.3.1 Consistency and Oracle Properties

Our first aim is to demonstrate that the SPECS estimator attains the same rate of convergence as the conventional least squares estimator.[4] Following standard convention in the cointegration literature, we first derive the consistency for a linear transformation of the coefficients to avoid singularities in the limits of sample moment matrices resulting from common stochastic trends (e.g. Lütkepohl, 2005, p. 290). In particular, under Assumption 3.2, the Granger Representation Theorem as displayed in Johansen (1995a, p. 49) enables (3.3) to be written as a VMA process of the form

$$\boldsymbol{z}_t = \boldsymbol{C}\boldsymbol{s}_t + \boldsymbol{\mu} + \boldsymbol{\tau}t + \boldsymbol{C}(L)\boldsymbol{\epsilon}_t + \boldsymbol{z}_0 = \boldsymbol{C}\boldsymbol{s}_t + \boldsymbol{\mu} + \boldsymbol{\tau}t + \boldsymbol{u}_t, \tag{3.11}$$

where $\boldsymbol{C} = \boldsymbol{B}_\perp \left(\boldsymbol{A}_\perp' \left(\boldsymbol{I}_N - \sum_{j=1}^{p} \boldsymbol{\Phi}_j \right) \boldsymbol{B}_\perp \right)^{-1} \boldsymbol{A}_\perp'$, $\boldsymbol{s}_t = \sum_{i=1}^{t} \boldsymbol{\epsilon}_i$, and $\boldsymbol{u}_t = \boldsymbol{C}(L)\boldsymbol{\epsilon}_t + \boldsymbol{z}_0$ a stationary process. In matrix notation, we write

$$\boldsymbol{Z}_{-1} = \boldsymbol{S}_{-1}C' + \boldsymbol{\iota}\boldsymbol{\mu}' + \bar{\boldsymbol{t}}\boldsymbol{\tau}' + \boldsymbol{U}, \tag{3.12}$$

with $\boldsymbol{S}_{-1} = (\boldsymbol{s}_0, \ldots, \boldsymbol{s}_{T-1})'$ and $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)'$. When cointegration is present in the data, the matrix $\boldsymbol{C}$ will be of rank $N - r$ such that the system may be separated into a stationary and non-stationary component. More specifically, we can define the linear transformation

$$\boldsymbol{Q} := \begin{bmatrix} \boldsymbol{B}' & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_M \\ \boldsymbol{A}_\perp' & \boldsymbol{0} \end{bmatrix} \text{ with } \boldsymbol{Q}^{-1} = \begin{bmatrix} \boldsymbol{A}(\boldsymbol{B}'\boldsymbol{A})^{-1} & \boldsymbol{0} & \boldsymbol{B}_\perp(\boldsymbol{A}_\perp'\boldsymbol{B}_\perp)^{-1} \\ \boldsymbol{0} & \boldsymbol{I}_M & \boldsymbol{0} \end{bmatrix}, \tag{3.13}$$

[4] As we derive our results for fixed N, we do not need to make an explicit assumption that the conditional model is sparse. Of course, in practical settings where T and N are of comparable size, sparsity is required for good performance. We return to this issue in our simulation study in Section 3.4.

such that we can decompose the system into

$$\boldsymbol{Q}\begin{bmatrix}\boldsymbol{z}_{t-1}\\ \boldsymbol{w}_t\end{bmatrix} = \begin{bmatrix}\boldsymbol{\xi}_{1,t}\\ \boldsymbol{\xi}_{2,t}\end{bmatrix}, \quad \boldsymbol{\xi}_{1,t} = \begin{bmatrix}\boldsymbol{B}'\boldsymbol{z}_{t-1}\\ \boldsymbol{w}_t\end{bmatrix}, \quad \text{and} \quad \boldsymbol{\xi}_{2,t} = \boldsymbol{A}_\perp'\boldsymbol{z}_{t-1}.$$

Then, $\boldsymbol{\xi}_{1,t}$ and $\boldsymbol{\xi}_{2,t}$ are a stationary and a non-stationary random vector, respectively.

Having defined the appropriate transformation, we are now able to state that SPECS attains the same rate of convergence as the OLS estimator. The proofs of all theorems in this section are provided in Appendix 3.A.2.

Theorem 3.1 (Estimation Consistency)**.** *Assume that*

$$\frac{\lambda_{G,T}\sigma_{Z,\max}}{\sqrt{T}} \xrightarrow{p} 0, \quad \frac{\lambda_{\delta,T}\sigma_{Z,\max}}{\sqrt{T}} \xrightarrow{p} 0 \quad \textit{and} \quad \frac{\lambda_{\pi,T}\sigma_{W,\max}}{\sqrt{T}} \xrightarrow{p} 0.$$

Let $\boldsymbol{D}_T = \operatorname{diag}(T\boldsymbol{I}_N, \sqrt{T}\boldsymbol{I}_M)$ *and* $\boldsymbol{S}_T = \operatorname{diag}(\sqrt{T}\boldsymbol{I}_{M+r}, T\boldsymbol{I}_{N-r})$*. Then, under Assumption 3.1 and 3.2, the estimators* $\hat{\boldsymbol{\gamma}}$ *satisfy:*

1. *No cointegration:* $\boldsymbol{D}_T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) = O_p(1)$*.*
2. *Cointegration:* $\boldsymbol{S}_T\boldsymbol{Q}'^{-1}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) = O_p(1)$*.*

The conditions imposed on the penalty terms limit the amount of shrinkage to prevent excessive shrinkage bias from impeding consistent estimation. Clearly, the admissible growth rates of the penalties are dependent on the stochastic order of the possibly random quantities $\sigma_{Z,\max}$ and $\sigma_{W,\max}$. Consequently, the practice of standardizing the data by scaling each variable by its corresponding estimated standard deviation may influence the restrictions imposed on the growth rate of the penalty. To illustrate, consider the case where $\boldsymbol{z}_t$ contains N random walks (with no drift components). Then, for any $i \in \{1, \dots, N\}$, the estimated standard deviation is

$$\hat{\sigma}_{Z,i} = \sqrt{\frac{\sum_{t=0}^{T-1} z_{it}^2}{T}} = \sqrt{T}\sqrt{\frac{\sum_{t=0}^{T-1} z_{it}^2}{T^2}} = O_p(\sqrt{T}),$$

such that also $\sigma_{Z,\max} = O_p(\sqrt{T})$. As a result, the requirements $\frac{\lambda_{G,T}\sigma_{Z,\max}}{\sqrt{T}} \xrightarrow{p} 0$ and $\frac{\lambda_{\delta,T}\sigma_{Z,\max}}{\sqrt{T}} \xrightarrow{p} 0$ translate to $\lambda_{G,T} \to 0$ and $\lambda_{\delta,T} \to 0$. While theoretically feasible, the notion of requiring a vanishing penalty to maintain consistent estimation does not conform with the belief of a sparse DGP. Moreover, the presence of deterministic components in the variables, such as a trend/drift, impact the stochastic order of the standard deviation and, hence, the required growth rates of the penalty. Therefore, we advise against the convention of standardization by the estimated standard deviations.

In situations where the data is clearly measured on drastically different scales, one may wish to apply an 'ad-hoc' standardization of the variables. As long as this standardization does not change the stochastic order of the data, the requirements on the amount of penalization remains the same and our theoretical results continue to go through. Possible choices therefore include to standardize by the standard deviations of first differences or AR(1) residuals, if theoretical guidance is not available (e.g. if variables are measured in different units, a logical standardization is often easy to find). Such choices result in standardizations that are, or converge to, constants, thereby not affecting the orders of the data, and allowing one to recover the original coefficients, if desired. Given the ad-hoc nature of such standardizations, the simulations and empirical application in this paper are conducted without standardization.

Remark 3.4. By construction of $\boldsymbol{Q}$, the resulting convergence stated in part (2) of Theorem 3.1 is equivalent to the statements $\boldsymbol{S}_T^* \boldsymbol{Q}^* (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) = O_p(1)$ and $\sqrt{T}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) = O_p(1)$, where $\boldsymbol{S}_T^* = \text{diag}(\sqrt{T}\boldsymbol{I}_r, T\boldsymbol{I}_{N-r})$ and $\boldsymbol{Q}^* = \begin{bmatrix} (\boldsymbol{A}'\boldsymbol{B})^{-1}\boldsymbol{A}' \\ (\boldsymbol{B}_\perp'\boldsymbol{A}_\perp)^{-1}\boldsymbol{B}_\perp \end{bmatrix}$.

SPECS performs continuous model selection by estimating sparse solutions through the imposition of individual L_1-penalties and a group penalty. In addition to consistently estimating the model parameters, an additional natural requirement of the estimator is to provide consistent selection of the relevant variables. This property is crucial when one aims to obtain interpretable solutions or even utilize the estimator as an alternative to classical tests for cointegration. An example of a traditional test for cointegration is the ECM-test by Banerjee et al. (1998) which looks at the t-ratio of the ordinary least squares coefficient of the lagged dependent variable. Alternatively, Boswijk (1994) proposes to test for the joint significance of the least squares coefficients of all lagged levels with a Wald-type test. One could interpret exclusion of the lagged levels of the dependent variable, or the lagged levels of all variables, as evidence against the presence of cointegration. However, an assumption of weak exogeneity is necessary when the aim is a direct interpretation of the estimated cointegration vector. Notwithstanding this caveat, selection consistency allows SPECS to be used as a screening mechanism that excludes irrelevant variables, even in the absence of weak exogeneity.[5]

Theorem 3.2 (Selection Consistency)**.** *Assume that*

$$\frac{\lambda_{\delta,T}\sigma_{Z,\min}}{T^{1-k_\delta/2}} \to \infty \text{ and } \frac{\lambda_{\pi,T}\sigma_{W,\min}}{T^{1/2-k_\pi/2}} \to \infty.$$

[5] A more detailed discussion of the interpretation of sparsity absent of weak exogeneity is provided in Section 3.3.2.

Then, under the same conditions as in Theorem 3.1, it holds that whenever $\gamma_i = 0$*, we have* $\mathbb{P}(\hat{\gamma}_i = 0) \to 1$.

Whereas the estimation consistency in Theorem 3.1 puts an upper limit on the amount of permissible shrinkage, the selection consistency in Theorem 3.2 requires a minimum amount of shrinkage to correctly remove irrelevant variables from the model. As before, the implied conditions regulating the growth rates of the penalties depend on the stochastic order of the possibly random quantities in $\boldsymbol{\Sigma}_V$. Assuming once more that $\boldsymbol{\Sigma}_Z$ is a diagonal matrix containing the standard deviations of the columns of $\boldsymbol{Z}_{-1}$, the condition for selection consistency of the lagged levels translates to $\frac{\lambda_{\delta,T}}{T^{1/2-k_\delta/2}} \to \infty$, as opposed to the $\frac{\lambda_{\delta,T}}{\sqrt{T}} \to 0$ required for estimation consistency. While any choice of $k_\delta > 0$ complies with these conditions from a theoretical point of view, we observe in simulations that the use of standard deviations as a means of standardization results in frequent removal of relevant non-stationary variables, thereby providing another argument against the use of standard deviations.

Remark 3.5. The only restriction imposed on the growth rate of the group penalty is $\frac{\lambda_{G,T}\sigma_{Z,\max}}{\sqrt{T}} \xrightarrow{p} 0$, which is necessary to avoid the shrinkage bias induced by the group penalty from impeding estimation consistency. Since $\lambda_{G,T} = 0$ is an admissible value, it follows that SPECS provides both consistent estimation and selection without the addition of a group penalty as well.

Remark 3.6. A common implementation of the adaptive lasso in the stationary setting sets $k_\delta = k_\pi = 1$. However, in the presence of cointegration the coefficients regulating the long-run dynamics are $\sqrt{T}$-consistent, whereas the presence of common stochastic trends demand a higher rate to stabilize the data. Consequently, assuming $\boldsymbol{\Sigma}_V = \boldsymbol{I}_{N+M}$, the conditions on λ_δ are $\frac{\lambda_\delta}{\sqrt{T}} \to 0$ and $\frac{\lambda_\delta}{T^{1-k_\delta/2}} \to \infty$. Hence, a choice of $k_\delta > 1$ is needed to maintain consistent selection of the lagged levels. Intuitively, one may argue that stricter penalization is necessitated by the correlation induced between the levels of variables through the presence of common stochastic trends.

Next, we establish that the limit distribution for the estimates of the non-zero population coefficients is the same as that of the oracle OLS estimator. When $\boldsymbol{\delta} \neq \boldsymbol{0}$, it follows from (3.11) that the subset of variables indexed by S_δ has the representation

$$\boldsymbol{z}_{S_\delta,t} = \boldsymbol{B}_{\perp,S_\delta} \left(\boldsymbol{A}_\perp' \left(\boldsymbol{I}_N - \sum_{j=1}^{p} \boldsymbol{\Phi}_j \right) \boldsymbol{B}_\perp \right)^{-1} \boldsymbol{A}_\perp \boldsymbol{s}_{t-1} + \boldsymbol{v}_{S_\delta,t}, \tag{3.14}$$

where $\boldsymbol{B}_{\perp,S_\delta}$ is a $(|S_\delta| \times (N-r))$-dimensional matrix. Let $\boldsymbol{B}^0_{S_\delta}$ denote the left

nullspace of $\boldsymbol{B}_{\perp,S_\delta}$, i.e.

$$\boldsymbol{B}^0_{S_\delta} = \left\{ \boldsymbol{x} \in \mathbb{R}^{|S_\delta|} | \boldsymbol{B}'_{\perp,S_\delta} \boldsymbol{x} = \boldsymbol{0} \right\}.$$

Note that by construction $\boldsymbol{B}'_{\perp,S_\delta} \boldsymbol{\delta}_{S_\delta} = \boldsymbol{0}$, such that $\dim(\boldsymbol{B}^0_{S_\delta}) = r_2 > 0$, where the dimension of the null space is defined as the number of linearly independent vectors in a corresponding basis.[6] For the case $|S_\delta| > r_2$, define $\boldsymbol{B}_{S_\delta}$ as a basis matrix, i.e. a $(|S_\delta| \times r_2)$-dimensional matrix whose columns form a basis for $\boldsymbol{B}^0_{S_\delta}$. Equivalently, define $\boldsymbol{B}_{S_\delta,\perp}$ as a $(|S_\delta| \times (|S_\delta| - r_2))$-dimensional basis matrix for the orthogonal complement of $\boldsymbol{B}_{S_\delta}$.[7] With the use of these linear transformations, we are able to confirm the convergence to the appropriate asymptotic distribution in the following theorem.

Theorem 3.3 (Limit Distribution)**.** *Define* $\boldsymbol{S}_{T,S_\gamma} = diag(\sqrt{T}\boldsymbol{I}_{|S_\pi|+r_2}, T\boldsymbol{I}_{|S_\delta|-r_2})$ *and*

$$\boldsymbol{Q}_{S_\gamma} = \begin{bmatrix} \boldsymbol{B}'_{S_\delta} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{|S_\pi|} \\ \boldsymbol{B}'_{S_\delta,\perp} & \boldsymbol{0} \end{bmatrix}, \quad \textit{such that}$$

$$\boldsymbol{Q}^{-1}_{S_\gamma} = \begin{bmatrix} \boldsymbol{B}_{S_\delta} \left(\boldsymbol{B}'_{S_\delta} \boldsymbol{B}_{S_\delta} \right)^{-1} & \boldsymbol{0} & \boldsymbol{B}_{S_\delta,\perp} \left(\boldsymbol{B}'_{S_\delta,\perp} \boldsymbol{B}_{S_\delta,\perp} \right)^{-1} \\ \boldsymbol{0} & \boldsymbol{I}_{|S_\pi|} & \boldsymbol{0} \end{bmatrix}.$$

Under the same assumptions as in Theorem 3.1 and 3.2 it holds that:

1. *No cointegration:* $\sqrt{T}(\hat{\boldsymbol{\pi}}_{S_\pi} - \hat{\boldsymbol{\pi}}_{OLS,S_\pi}) = o_p(1)$.
2. *Cointegration:* $\boldsymbol{S}_{T,S_\gamma} \boldsymbol{Q}'^{-1}_{S_\gamma} (\hat{\boldsymbol{\gamma}}_{S_\gamma} - \hat{\boldsymbol{\gamma}}_{OLS,S_\gamma}) = o_p(1)$.

Remark 3.7. When all variables in $\boldsymbol{z}_{S_\delta,t}$ are stationary, it must hold that $\boldsymbol{B}_{\perp,S_\delta} = \boldsymbol{0}$ such that $r_2 = \dim(\boldsymbol{B}^0_{S_\delta}) = |S_\delta|$. In this special case we define $\boldsymbol{Q}_{S_\gamma} = \boldsymbol{I}_{|S_\gamma|}$ and $\boldsymbol{S}_{T,S_\gamma} = \sqrt{T}\boldsymbol{I}_{s_\gamma}$.

As a direct consequence of Theorem 3.3, we obtain the limit distribution of the SPECS estimator scaled by $\sqrt{T}$.

Corollary 3.1. *Under the same conditions as in Theorem 3.3, we have*

$$\sqrt{T} \left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma} \right) \xrightarrow{d} \mathcal{N} \left(\boldsymbol{0}, \begin{bmatrix} \boldsymbol{B}_{S_\delta} \boldsymbol{\Sigma}^{-1}_U \boldsymbol{B}'_{S_\delta} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{W_{S_\pi}} \end{bmatrix} \right), \tag{3.15}$$

[6] For details on the existence of a basis and its relation to the dimension of a finite-dimensional vector space, see Abadir and Magnus (2005, ex. 3.25, 3.29 and 3.30).

[7] Hence, $\boldsymbol{B}_{\perp,S_\delta}$ are the rows of $\boldsymbol{B}_\perp$ indexed by S_δ, whereas $\boldsymbol{B}_{S_\delta,\perp}$ is a matrix whose columns form a basis for the orthogonal complement of $\boldsymbol{B}_{S_\delta}$.

where $\boldsymbol{\Sigma}_U = \mathbb{E}\left(\boldsymbol{B}'_{S_\delta}\boldsymbol{u}_{S_\delta,t}\boldsymbol{u}'_{S_\delta,t}\boldsymbol{B}_{S_\delta}\right)$ *and* $\boldsymbol{\Sigma}_{W_{S_\pi}} = \mathbb{E}\left(\boldsymbol{w}_{S_\pi,t}\boldsymbol{w}'_{S_\pi,t}\right)$. *Furthermore, the matrix* $\boldsymbol{B}_{S_\delta}\boldsymbol{\Sigma}_U^{-1}\boldsymbol{B}'_{S_\delta}$ *is uniquely defined regardless of the choice of basis matrix* $\boldsymbol{B}_{S_\delta}$.

Remark 3.8. The oracle results in Theorem 3 suggest that one could test for cointegration by applying standard low-dimensional cointegration tests, such as the Wald test by Boswijk (1994), on the selected variables with the same asymptotic distribution as if only the selected variables were considered from the start. However, such a post-selection inferential procedure should be treated with caution, as it is well known that the selection step impacts the sampling properties of the estimator (see Leeb and Pötscher, 2005). The convergence results of many selection procedures, SPECS included, hold pointwise only, with the resulting implication that the finite-sample distribution will not get uniformly close to the respective asymptotic distribution when the sample size grows large. The practical implication is that for certain values of the parameters in the underlying DGP, relying on the oracle properties for post-selection test statistics may be misleading. Developing a valid post-selection cointegration test is certainly of interest. However, the field of valid post-selection inference is, while rapidly developing, still in its infancy. None of the currently existing methods, such as those considered in Berk et al. (2013), Van de Geer et al. (2014), Lee et al. (2016) or Chernozhukov et al. (2018), can easily be adapted to - let alone validated in - our setting. Developing such a method therefore requires a full new theory which is outside the scope of the current chapter.

Finally, all results thus far have focussed on the convergence and selection of the coefficients corresponding to the stochastic component in our model. Based on these results, we are able to obtain the behaviour of the estimated coefficients governing the deterministic components. However, the rate of convergence of the trend coefficient depends on three characteristics of the DGP, namely the presence of cointegration, the presence of a deterministic trend and whether the trend occurs within the long-run equilibrium. Consequently, we state the following corollary, the proof of which is delegated to the supplementary appendix.

Corollary 3.2. *Under the assumptions in Theorem 3.1 and 3.2, the estimators of the coefficients regulating the deterministic component, i.e.* $\hat{\mu}_0$ *and* $\hat{\tau}_0$*, are consistent. In particular, we have*

$$\sqrt{T}(\hat{\mu}_0 - \hat{\mu}_{0,OLS}) = o_p(1),$$
$$R_T(\hat{\tau}_0 - \hat{\tau}_{0,OLS}) = o_p(1),$$

$$\textit{where } R_T = \begin{cases} T^{3/2} & \boldsymbol{\tau} = 0 \\ T & \boldsymbol{\tau} \neq 0, \boldsymbol{B}'\tau = 0 \\ T^{1/2} & \boldsymbol{\tau} \neq 0, \boldsymbol{B}'\boldsymbol{\tau} \neq 0 \end{cases}.$$

In summary, under appropriate assumptions on the penalty rates, SPECS is able to consistently estimate the coefficients of the relevant stochastic variables with the same rate and asymptotic efficiency as the oracle least squares estimator and the inclusion of unpenalized deterministic components allows for an invariant limiting distribution in the same way de-meaning and de-trending is performed in the least squares case. In addition, the irrelevant variables are removed from the model with probability approaching one.

Remark 3.9. A possible extension to consider is allowing SPECS to select the appropriate deterministic specification by penalizing the coefficients corresponding to a set of deterministic components. While this certainly would be straightforward to implement, the extension of the current theoretical results to this new estimator is less trivial for two main reasons. The first difficulty is that the presence of a trend or drift component in a variable dominates its stochastic variation asymptotically, such that appropriately scaled estimates of sample covariance matrices converge to reduced rank matrices. This feature becomes problematic in instances where inverses or positive minimum eigenvalues are required. While the inclusion of unpenalized deterministic components allows one to effectively regress out the effect of those components (Yamada, 2017), this is not the case when the deterministic components are penalized as well. Secondly, the (pointwise) asymptotic distributions of the estimators are not uniquely identified when the trend coefficient is penalized. Based on the definition given in (3.9), a specification where $\tau_0 = 0$ can be implied by either (i) $\boldsymbol{\tau} = \mathbf{0}$ or (ii) $\boldsymbol{\tau} \neq \mathbf{0}$ and $\boldsymbol{\delta}'\boldsymbol{\tau} = 0$. It is well known that the limit distribution varies depending on whether a deterministic trend is present in the data (Park and Phillips, 1988, Theorems 3.2 and 3.3), such that identification of the distribution is not ensured when the data is not first de-trended.

3.3.2 Implications for Particular Model Specifications

To fully appreciate the theoretical results in the preceding section, a detailed understanding of the generality provided by the set of imposed assumptions is helpful. For example, as the results are derived without requiring weak exogeneity, our set of assumptions allows for the presence of stationary variables in the data. However, in the absence of weak exogeneity, model interpretation becomes non-standard. Therefore, in this section we elaborate on several relevant model specifications to demonstrate

the flexibility of the single-equation model and highlight the practical implications of variable selection in such a general framework.

Mixed Orders of Integration

One of the most prominent benefits of SPECS is the ability to model potentially non-stationary and cointegrated data without the need to adopt a pre-testing procedure with the aim of checking, and potentially correcting, for the order of integration or to decide on the appropriate cointegrating rank of the system. Assumptions 3.1 and 3.2 under which our theory is developed are compatible with a wide variety of DGPs that include settings where the dataset contains an arbitrary mix of $I(1)$ and $I(0)$ variables. The dataset is simply transformed according to (3.7) and SPECS provides consistent estimation of the parameters and consistently identifies the correct implied sparsity pattern. The purpose of this section is to demonstrate this feature by means of some illustrative examples.

The central idea underlying the above feature is that a single-equation model can be derived from any system admitting a finite order VECM representation. In a VECM system containing variables with mixed orders of integration, each stationary variable adds an additional trivial cointegrating vector. Such a vector corresponds to a basis vector that equals one on the index of the stationary variable. For illustrative purposes, we consider the following general example. Define $\boldsymbol{z}_t = (\boldsymbol{z}'_{1,t}, \boldsymbol{z}'_{2,t})'$, where $\boldsymbol{z}_{1,t} \sim I(0)$ and $\boldsymbol{z}_{2,t} \sim I(1)$ and possibly cointegrated. Let the dimensions of $\boldsymbol{z}_{1,t}$ and $\boldsymbol{z}_{2,t}$ be N_1 and N_2 respectively. Then, $\boldsymbol{z}_t$ admits the representation

$$\begin{aligned}\begin{bmatrix}\Delta \boldsymbol{z}_{1,t} \\ \Delta \boldsymbol{z}_{2,t}\end{bmatrix} &= \begin{bmatrix}-\boldsymbol{I}_{N_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}\end{bmatrix}\begin{bmatrix}\boldsymbol{z}_{1,t-1} \\ \boldsymbol{z}_{2,t-2}\end{bmatrix} + \boldsymbol{\Phi}(L)\Delta \boldsymbol{z}_{t-1} + \boldsymbol{\epsilon}_t \\ &= \boldsymbol{B}\boldsymbol{z}_{t-1} + \boldsymbol{\Phi}(L)\Delta \boldsymbol{z}_{t-1} + \boldsymbol{\epsilon}_t,\end{aligned}$$

where $\boldsymbol{\Phi}(L)$ corresponds to a p-dimensional matrix lag polynomial by Assumption 3.2 and $\boldsymbol{\epsilon}_t$ satisfies the conditions in Assumption 3.1. In addition, we maintain the convention that $\boldsymbol{A} = \boldsymbol{0}$ when $\boldsymbol{z}_{2,t}$ does not cointegrate. Naturally, the single-equation derived from this VECM has the same form as in (3.7), with the crucial difference that some of the variables in $\boldsymbol{z}_{t-1}$ are stationary. More specifically, let $\boldsymbol{\pi}_0$ be defined as in (3.6) with the decomposition $\boldsymbol{\pi}_0 = (\boldsymbol{\pi}'_{0,1}, \boldsymbol{\pi}'_{0,2})'$. Without loss of generality, if $y_t \sim I(0)$ we let $\boldsymbol{z}_{1,t} = (y_t, \boldsymbol{x}'_{1,t})'$, whereas if $y_t \sim I(1)$ we let $\boldsymbol{z}_{2,t} = (y_t, \boldsymbol{x}'_{2,t})'$. The

single-equation model can then be represented as usual

$$\begin{aligned}\Delta y_t &= (1, -\boldsymbol{\pi}_0') \left(\boldsymbol{B}\boldsymbol{z}_{t-1} + \boldsymbol{\Phi}(L)\Delta\boldsymbol{z}_{t-1}\right) + \boldsymbol{\pi}_0'\Delta\boldsymbol{x}_t + \epsilon_{y,t} \\ &= \delta'\boldsymbol{z}_{t-1} + \boldsymbol{\pi}'\boldsymbol{w}_t + \epsilon_{y,t}.\end{aligned} \tag{3.16}$$

or alternatively

$$\Delta y_t = \boldsymbol{\delta}_2'\boldsymbol{z}_{2,t-1} + \boldsymbol{\pi}^{*\prime}\boldsymbol{w}_t^* + \epsilon_{y,t}, \tag{3.17}$$

where $\boldsymbol{\pi}^* = (\boldsymbol{\delta}_1', \boldsymbol{\pi}')'$ and $\boldsymbol{w}_t^* = (\boldsymbol{z}_{1,t-1}', \boldsymbol{w}_t')'$. This representation highlights that the single-equation model can be decomposed into contributions from the non-stationary variables, i.e. $\boldsymbol{z}_{2,t-1}$, and stationary variables, i.e. $\boldsymbol{w}_t^*$. Moreover, from our theoretical results in Theorem 3.3 it follows that

$$\sqrt{T}(\hat{\boldsymbol{\delta}}_1 - \hat{\boldsymbol{\delta}}_{1,OLS}) = o_p(1). \tag{3.18}$$

In the extreme case, where the DGP consists of a collection of stationary variables and a collection of variables that are integrated of order one which do not cointegrate, we have $\boldsymbol{B}_{\perp,S_\delta} = \boldsymbol{0}$ such that (3.18) follows directly from Remark 3.7.

Finally, in Assumption 3.2 we allow for the case where $\text{rank}(\boldsymbol{B}) = N$. One, perhaps slightly cumbersome, interpretation of this scenario is a system in which every variable 'trivially cointegrates', which intuitively motivates the applicability of our theoretical results. However, a more common interpretation follows from noting that when $r = N$ the system can be appropriately described by a stationary vector autoregressive model of the form

$$\boldsymbol{z}_t = \boldsymbol{\Phi}(L)\boldsymbol{z}_{t-1} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{\epsilon}_t$ complies with Assumption 3.1 and $\boldsymbol{\Phi}(L)$ denotes an invertible matrix lag-polynomial of order p. Following the procedure detailed in section 3.2, the corresponding single-equation model can be derived as

$$\begin{aligned}y_t &= \boldsymbol{\pi}'\boldsymbol{x}_t + (1, -\boldsymbol{\pi}')\boldsymbol{\Phi}(L)\boldsymbol{z}_{t-1} + \epsilon_{y,t} \\ &= \boldsymbol{\pi}'\boldsymbol{x}_t + (1, -\boldsymbol{\pi}')\boldsymbol{\Phi}(1)\boldsymbol{z}_{t-1} + (1, -\boldsymbol{\pi}')\tilde{\boldsymbol{\Phi}}(L)\Delta\boldsymbol{z}_{t-1} + \epsilon_{y,t},\end{aligned} \tag{3.19}$$

where the second equation follows from applying the Beveridge-Nelson decomposition

to $\boldsymbol{\Phi}(L)$. We can rewrite (3.19) as

$$\begin{aligned}\Delta y_t &= -y_{t-1} + \boldsymbol{\pi}'\boldsymbol{x}_{t-1} + \boldsymbol{\pi}'\Delta\boldsymbol{x}_t + (1, -\boldsymbol{\pi}')\boldsymbol{\Phi}(1)\boldsymbol{z}_{t-1} \\ &\quad + (1, -\boldsymbol{\pi}')\tilde{\boldsymbol{\Phi}}(L)\Delta\boldsymbol{z}_{t-1} + \epsilon_{y,t} \\ &= \boldsymbol{\delta}'\boldsymbol{z}_{t-1} + \boldsymbol{\pi}'\Delta\boldsymbol{x}_t + \boldsymbol{\Phi}^*(L)\Delta\boldsymbol{z}_{t-1} + \epsilon_{y,t},\end{aligned} \tag{3.20}$$

where $\boldsymbol{\delta} = (1, -\boldsymbol{\pi}')\left(-\boldsymbol{I} + \boldsymbol{\Phi}(1)\right)$ and $\boldsymbol{\Phi}^*(L) = (1, -\boldsymbol{\pi}')\tilde{\boldsymbol{\Phi}}(L)$. Hence, the single-equation model that we estimate can be derived from a stationary system as well. Given that all variables in (3.20) are stationary time series, SPECS can also be shown to consistently estimate the parameters based on the well-documented properties of the adaptive lasso in stationary time series settings, such as those considered in Medeiros and Mendes (2016).

Sparsity and Weak Exogeneity

The benefit of L_1-regularized estimation stems from its ability to identify sparse parameter structures. However, the concept of sparsity in the conditional model here considered merits additional clarification, as the potential absence of weak exogeneity obscures standard interpretability. In Section 3.2 we argue that the coefficients regulating the long-run dynamics in the conditional model are generally derived from linear combinations of the cointegrating vectors in the VECM representation (3.3). By decomposing the matrix with adjustment rates as $\boldsymbol{A} = \begin{bmatrix}\boldsymbol{\alpha}_1' \\ \boldsymbol{A}_2'\end{bmatrix}$, with $\boldsymbol{\alpha}_1$ an r-dimensional column-vector, we obtain

$$\boldsymbol{\delta} = \boldsymbol{B}(\boldsymbol{\alpha}_1 - \boldsymbol{A}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}).$$

It follows that $\delta_i = 0$ if the condition

$$\boldsymbol{\beta}_i'\left(\boldsymbol{\alpha}_1 - \boldsymbol{A}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}\right) = \boldsymbol{0} \tag{3.21}$$

is satisfied, where $\boldsymbol{\beta}_i$ is the i-th row-vector of $\boldsymbol{\beta}$. While this condition may hold in a variety of non-trivial ways, some general cases can be derived that lead to sparsity in $\boldsymbol{\delta}$. For example, a variable $\boldsymbol{x}_i$ that does not cointegrate with any of the variables in the system, i.e. $\boldsymbol{\beta}_i = \boldsymbol{0}$, will carry a zero coefficient in the derived long-run equilibrium in the single-equation model.

An additional special case is the addition of I(0) variables to the system. Consider the estimation of a standard VECM of the form (3.3) without any short-run dynamics. Assume, however, that the last variable in the dataset, say $z_{N,t}$, is a stationary white

noise series that is mistakenly considered to be integrated of order one. Denote the non-stationary variables by $\boldsymbol{z}_{1,t} = (z_{1,t}, \dots, z_{N-1,t})'$. Then, the simple VECM without short-run dynamics is described by the representation

$$\begin{bmatrix} \Delta \boldsymbol{z}_{1,t} \\ \Delta z_{N,t} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{0} \\ \boldsymbol{0} & -1 \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_1' & \boldsymbol{0} \\ \boldsymbol{0} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_{1,t-1} \\ z_{N,t-1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_{1,t} \\ \epsilon_{N,t} \end{bmatrix} = \boldsymbol{A}\boldsymbol{B}'\boldsymbol{z}_{t-1} + \boldsymbol{\epsilon}_t.$$

Letting the last row-vector of $\boldsymbol{B}$ be denoted by $\boldsymbol{\beta}_N = (0, \dots, 0, 1)'$, condition (3.21) then translates to $\boldsymbol{\beta}_N' \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_{21} = \boldsymbol{0}$. A sufficient condition for this to hold is when $\mathbb{E}(\epsilon_{N,t}\boldsymbol{\epsilon}_{1,t}) = 0$, implying that exogenous stationary variables will not be considered as part of the cointegration vector $\boldsymbol{\delta}$. This statement does not come at a surprise, but it also highlights that stationary variables whose errors are correlated with other variables in the system might end up being part of the cointegration vector in the equation for Δy_t. As this correlation contains information about Δy_t, we consider this property desirable for applications such as nowcasting. It does, however, demonstrate that care has to be taken when the aim is direct interpretation of the implied cointegrating vector in the absence of weak exogeneity.

Finally, we explore a somewhat less trivial case by considering a VECM model in which $\boldsymbol{\Sigma}$, the covariance matrix of the errors, follow a Toeplitz structure with $\sigma_{ij} = \rho^{|i-j|}$. After partitioning $\boldsymbol{\Sigma}$ as in (3.5), we can rewrite

$$\boldsymbol{\sigma}_{21} = \begin{bmatrix} \rho^1 \\ \vdots \\ \rho^N \end{bmatrix} = \begin{bmatrix} \rho^0 & \dots & \rho^{N-1} \\ \vdots & \ddots & \vdots \\ \rho^{N-1} & \dots & \rho^0 \end{bmatrix} \begin{bmatrix} \rho^1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \boldsymbol{\Sigma}_{22}\boldsymbol{\pi}_0, \tag{3.22}$$

thus showing that $\boldsymbol{\pi}_0 = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21} = (\rho, 0, \dots, 0)'$.[8] As $\boldsymbol{\delta}' = (1, -\boldsymbol{\pi}_0')\boldsymbol{A}\boldsymbol{B}'$, this implies that only the long-run equilibria that occur in the equations for Δy_t or its cross-sectionally neighbouring variable will be part of the linear combination in the derived the single-equation model. Consequently, any variables in the dataset that are not contained in the equilibria occurring in these equations will induce sparsity in $\boldsymbol{\delta}$.

[8] It is straightforward to show that this property carries over to covariance matrices with a block-diagonal Toeplitz structure, with each block $\boldsymbol{\Sigma}^{(k)}$ having the form $\sigma_{i,j}^{(k)} = \rho_{(k)}^{|i-j|}$. The number of non-zero elements in the resulting vector $\boldsymbol{\pi}_0$ will equal the number of blocks in the covariance matrix.

3.4 Simulations

In this section we analyze the selective capabilities and predictive performance of SPECS by means of simulations. We estimate the single-equation model according to the objective function (3.9) with the following settings for the penalty rates:

1. Ordinary Least Squares (OLS: $\lambda_{G,T} = 0$, $\lambda_{\delta,T} = 0$, $\lambda_{\pi,T} = 0$),
2. Autoregressive Distributed Lag (ADL: $\lambda_{G,T} = 0$, $\lambda_{\delta,T} = \infty$, $\lambda_{\pi,T} > 0$),
3. SPECS - no group penalty (SPECS$_1$: $\lambda_{G,T} = 0$, $\lambda_{\delta,T} > 0$, $\lambda_{\pi,T} > 0$),
4. SPECS - group penalty (SPECS$_2$: $\lambda_{G,T} > 0$, $\lambda_{\delta,T} > 0$, $\lambda_{\pi,T} > 0$).[9]

The OLS estimator is only included when feasible according to the dimension of the model to estimate and we additionally include a penalized autoregressive distributed lag model (ADL) with all variables entering in first differences. The latter model can be interpreted as the conditional model one would obtain when ignoring cointegration in the data and specifying a VAR in differences as a model for the full system. The resulting conditional model is the same as the CECM that we consider, but with the built-in restriction $\boldsymbol{\delta} = \mathbf{0}$.

For the sake of computational efficiency we estimate the solutions for $\lambda_{\delta,T}$ and $\lambda_{\pi,T}$ over a one-dimensional grid, i.e. both penalties are governed by a single universal parameter $\lambda_{I,T}$. We weigh the universal parameter by initial estimates obtained from a ridge regression. Specifically, we adopt $\omega_{\delta,i}^{k_\delta} = 1/\left|\hat{\delta}_{ridge,i}\right|^{k_\delta}$ and $\omega_{\pi,j}^{k_\pi} = 1/|\hat{\pi}_{ridge,j}|^{k_\pi}$, where $k_\delta = 2$ and $k_\pi = 1$ in accordance with the assumptions in Theorems 3.1 and 3.2. We consider 100 possible values for $\lambda_{I,T}$ and choose the final model based on the BIC criterion. For SPECS$_2$, the model selection takes place over a two-dimensional grid consisting of 100 values for $\lambda_{I,T}$ and 10 possible values for $\lambda_{G,T}$. We note that while the use of the single universal penalty $\lambda_{I,T}$ significantly reduces the dimension of the search space, this heuristic may negatively impact the performance of SPECS. Since this choice of implementation does not impact the ADL model, the relative performance gain of SPECS over the ADL model would likely be underestimated.

We now consider three different settings under which we analyze the performance of our SPECS estimator.

[9] As a useful mnemonic, the reader may relate the subscript to the number of penalty categories included in the estimation; SPECS$_1$ only contains an individual penalty whereas SPECS$_2$ contains both a group penalty and individual penalty.

Table 3.1 Simulation Design for the First Study (Dimensionality and Weak Exogeneity)

Low Dimension	$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{\delta}$
WE	$\begin{bmatrix} \alpha_1 \\ \mathbf{0}_{9\times 1} \end{bmatrix}$	$\begin{bmatrix} \tilde{\boldsymbol{\iota}} \\ \mathbf{0}_{5\times 1} \end{bmatrix}$	$\alpha_1 \boldsymbol{B}$
No WE	$\alpha_1 \boldsymbol{B}$	$\begin{bmatrix} \tilde{\boldsymbol{\iota}} & \mathbf{0}_{5\times 1} \\ \mathbf{0}_{5\times 1} & \tilde{\boldsymbol{\iota}} \end{bmatrix}$	$\begin{bmatrix} (1+\rho)\alpha_1 \tilde{\boldsymbol{\iota}} \\ \mathbf{0}_{5\times 1} \end{bmatrix}$
High Dimension	$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{\delta}$
WE	$\begin{bmatrix} \alpha_1 \\ \mathbf{0}_{49\times 1} \end{bmatrix}$	$\begin{bmatrix} \tilde{\boldsymbol{\iota}} \\ \mathbf{0}_{45\times 1} \end{bmatrix}$	$\alpha_1 \boldsymbol{B}$
No WE	$\alpha_1 \boldsymbol{B}$	$\begin{bmatrix} \tilde{\boldsymbol{\iota}} & \mathbf{0}_{5\times 1} & \mathbf{0}_{5\times 1} \\ \mathbf{0}_{5\times 1} & \tilde{\boldsymbol{\iota}} & \mathbf{0}_{5\times 1} \\ \mathbf{0}_{5\times 1} & \mathbf{0}_{5\times 1} & \tilde{\boldsymbol{\iota}} \\ \mathbf{0}_{35\times 1} & \mathbf{0}_{35\times 1} & \mathbf{0}_{35\times 1} \end{bmatrix}$	$\begin{bmatrix} (1+\rho)\alpha_1 \tilde{\boldsymbol{\iota}} \\ \mathbf{0}_{45\times 1} \end{bmatrix}$

Notes: The low-dimensional (high-dimensional) design corresponds to a system with $N = 10$ ($N = 50$) unique time series and $N' = 31$ ($N' = 151$) parameters to estimate. Furthermore, $\tilde{\boldsymbol{\iota}} = (1, -\boldsymbol{\iota}_4')'$ and $\alpha_1 = -0.5, -0.45, \ldots, 0$ regulates the adjustment rate towards the equilibrium.

3.4.1 Dimensionality and Weak Exogeneity

In the first part of our simulation study we focus on the effects of dimensionality and weak exogeneity on a (co)integrated dataset. The general DGP from which we simulate our data is given by the equation

$$\Delta \boldsymbol{z}_t = \boldsymbol{A}\boldsymbol{B}'\boldsymbol{z}_{t-1} + \boldsymbol{\Phi}_1 \Delta \boldsymbol{z}_{t-1} + \boldsymbol{\epsilon}_t, \tag{3.23}$$

with $t = 1, \ldots, T = 100$, and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ with $\sigma_{ij} = 0.8^{|i-j|}$. Furthermore, $\boldsymbol{\Phi}_1$, the coefficient matrix regulating the short-run dynamics is generated as $0.4 \cdot \boldsymbol{I}_N$, where N varies depending on the specific DGP considered. Based on this DGP, the single-equation model takes on the form

$$\Delta y_t = \boldsymbol{\delta}' \boldsymbol{z}_{t-1} + \boldsymbol{\pi}_0' \Delta \boldsymbol{x}_t + \boldsymbol{\pi}_1' \Delta \boldsymbol{z}_{t-1} + \epsilon_{y,t},$$

with $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ as defined in (3.7). We consider a total of four different settings, corresponding to different combinations of (i) dimensionality (low/high) and (ii) weak exogeneity (present/absent). The corresponding parameter settings, and their implied cointegrating vector $\boldsymbol{\delta}$, are tabulated in Table 3.1.

We measure the selective capabilities based on three metrics. The pseudo-power of the models measures the ability to appropriately pick up the presence of cointegration in the underlying DGP. For the OLS procedure we perform the Wald test proposed

by Boswijk (1994). When the OLS fitting procedure is unfeasible due to the high-dimensionality, we perform the Wald test on the subset of variables included after fitting SPECS_1 and refer to this approach as Wald-PS (where PS stands for post-selection). Despite the caveats of oracle-based post-selection inference mentioned in Remark 3.8, the inclusion of Wald-PS still offers valuable insights regarding the performance one may expect of such a procedure in light of the aforementioned limitation. SPECS is used as an alternative to this cointegration test by simply checking whether at least one of the lagged levels is included in the model. The percentage of trials in which cointegration is found is then reported as the pseudo-power.

Second, for each trial the Proportion of Correct Selection (PCS) describes the proportion of correctly selected variables:

$$PCS = \frac{|\{j : \hat{\gamma}_j \neq 0 \text{ and } \gamma_j \neq 0\}|}{|\{j : \gamma_j \neq 0\}|},$$

where $|\cdot|$ denotes the cardinality. Alternatively, the Proportion of Incorrect Selection (PICS) describes, as the name may suggest, the proportion of incorrectly selected variables:

$$PICS = \frac{|\{j : \hat{\gamma}_j \neq 0 \text{ and } \gamma_j = 0\}|}{|\{j : \gamma_j = 0\}|}.$$

The PCS and PICS are calculated for SPECS_1 and SPECS_2 and averaged over all trials.

Finally, we consider the predictive performance in a simulated nowcasting application, where we implicitly assume that the information on the latest realization of $\boldsymbol{x}_T$ arrives before the realization of y_T. These situations frequently occur in practice, see Giannone et al. (2008) and the references therein for an overview as well as the empirical application considered in Section 3.5. Due to the construction of the single-equation model, in which contemporaneous values of the conditioning variables contribute to the contemporaneous variation in the dependent variable, our proposed method is particularly well-suited to this application. For any of the considered fitting procedures, the nowcast is given by $\hat{y}_T = \hat{\boldsymbol{\delta}}' \boldsymbol{z}_{T-1} + \hat{\boldsymbol{\pi}}' \Delta \boldsymbol{x}_T + \hat{\boldsymbol{\phi}}' \Delta \boldsymbol{z}_{T-1}$, where by construction $\hat{\boldsymbol{\delta}} = \mathbf{0}$ in the ADL model. For each method we record the root mean squared nowcast error (RMSNE) relative to the OLS oracle procedure fitted on the subset of relevant variables.

Figure 3.1 visually displays the evolution of our performance metrics over a range of values for α_1, representing increasingly faster rates of adjustment towards the long-

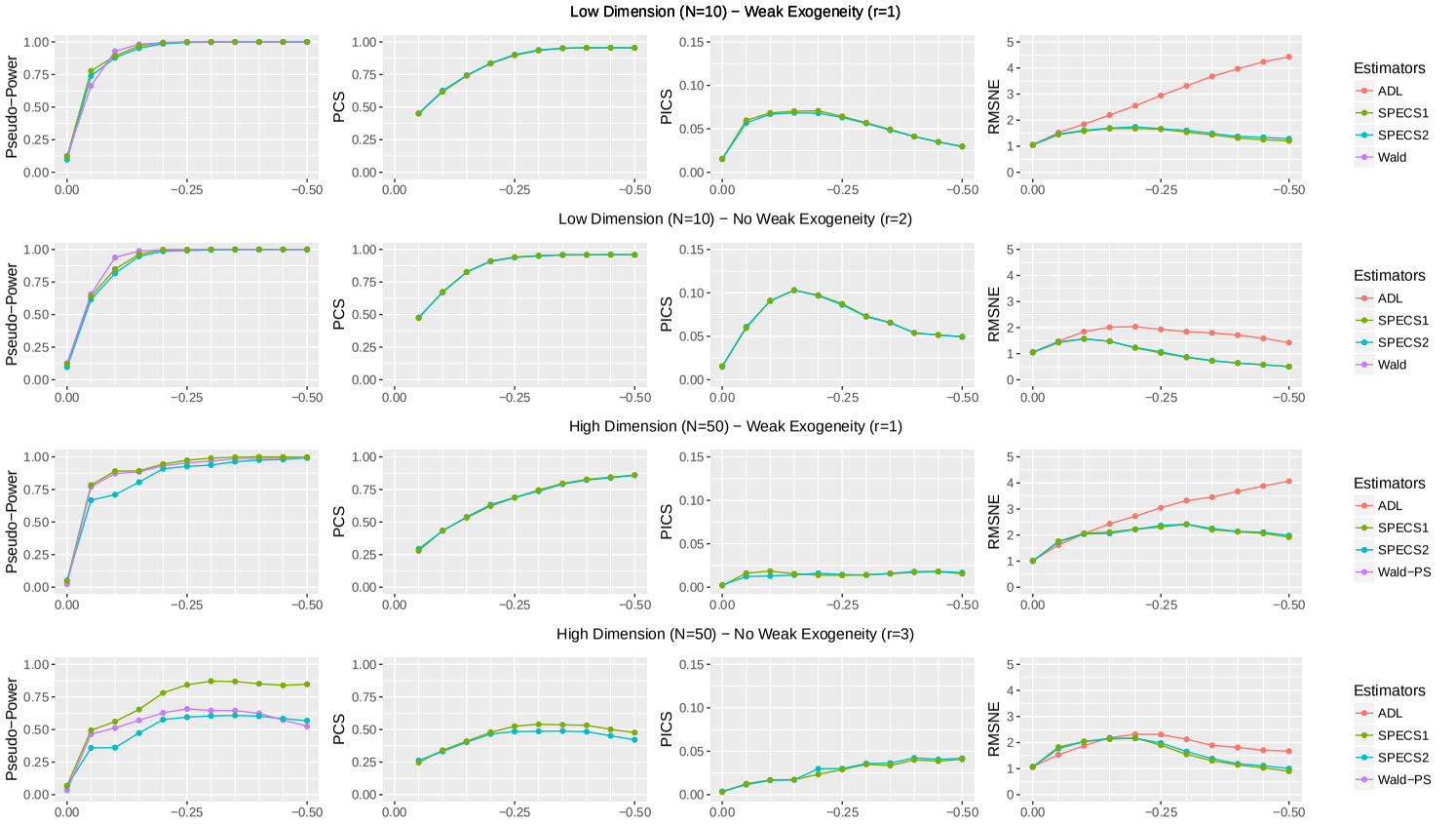

Figure 3.1: Pseudo-Power, Proportion of Correct Selection (PCS), Proportion of Incorrect Selection (PICS) and Root Mean Squared Nowcast Error (RMSNE) for Low- and High-Dimensional specifications. The adjustment rate multiplier α_1 is on the horizontal axis.

run equilibrium. The first row of plots shows near-perfect performance of SPECS over all metrics. The pseudo-size is slightly lower than the size of the Wald test when the latter is controlled at 5%, whereas the pseudo-power quickly approaches one. Following expectations, the pseudo-size for SPECS_2 is slightly lower as a result of the additional group penalty. Focussing on the selection of variables, we find that for faster adjustment rates, SPECS is able to exactly identify the sparsity pattern with very high frequency, as demonstrated by the PCS approaching 100% and the PICS staying near 0%. Furthermore, the MSNE obtained by our methods is close to the oracle method and is substantially lower than the MSNE obtained by the ADL model for faster adjustment rates, while being almost identical absent of cointegration. The picture remains qualitatively similar when moving away from weak exogeneity while staying in a low-dimensional framework, although the gain in predictive performance over the ADL decreases somewhat. We postulate that the ADL may benefit from a bias-variance tradeoff, especially considering that the correctly specified single-equation model is sub-optimal in terms of efficiency absent of weak exogeneity compared to a full system estimator. Nonetheless, SPECS is clearly the preferred method.

The performance in the high-dimensional setting is displayed in rows 3 and 4 of Figure 3.1. When the conditioning variables are weakly exogenous with respect to the parameters of interest, the selective capabilities remain strong. The pseudo-power demonstrates the attractive prospect of using our method as an alternative to cointegration testing, especially when taking into consideration that the traditional Wald test is infeasible in the current setting. In addition, the nowcasting performance remains far superior to that of the misspecified ADL. The last row depicts the performance absent of weak exogeneity. In this setting, exact identification of the implied cointegrating vector occurs less frequently, which seems to negatively impact the nowcasting performance. However, the misspecified ADL is still outperformed, despite the deterioration in the selective capabilities of our method.

3.4.2 Mixed Orders of Integration

We move on to an analysis of the performance of SPECS on datasets containing variables with mixed orders of integration. The aim of this section is to gain an understanding of the relative performance of SPECS when not all time series are (co)integrated and to compare the performance of SPECS to traditional approaches that rely on pre-testing. The latter goal is attained by adding an additional penalized ADL model to the comparison, namely one in which the data is first corrected for non-stationarity based on a pre-testing procedure in which an Augmented Dickey-Fuller (ADF) test is performed on the individual series. We refer to this procedure as

Table 3.2 Simulation Design for the Second Study: Mixed Orders of Integration

Order	$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{\delta}$
$y \sim I(0)$	$\begin{bmatrix} 1 & 0 & \mathbf{0}_{1\times 24} \\ \mathbf{0}_{15\times 1} & \alpha_1 \boldsymbol{B}^* & \mathbf{0}_{15\times 24} \\ \mathbf{0}_{10\times 1} & \mathbf{0}_{10\times 3} & \mathbf{0}_{10\times 24} \\ \mathbf{0}_{24\times 1} & \mathbf{0}_{24\times 3} & \boldsymbol{I}_{24} \end{bmatrix}$	$\begin{bmatrix} -b & 0 & \mathbf{0}_{1\times 24} \\ \mathbf{0}_{15\times 1} & \boldsymbol{B}^* & \mathbf{0}_{15\times 24} \\ \mathbf{0}_{10\times 1} & \mathbf{0}_{10\times 3} & \mathbf{0}_{10\times 24} \\ \mathbf{0}_{24\times 1} & \mathbf{0}_{24\times 3} & -\boldsymbol{B}_{24\times 24} \end{bmatrix}$	$\begin{bmatrix} -1 \\ -\rho\alpha_1 \tilde{\boldsymbol{\iota}} \\ \mathbf{0}_{44\times 1} \end{bmatrix}$
$y \sim I(1)$	$\begin{bmatrix} \alpha_1 \boldsymbol{B}^* & \mathbf{0}_{15\times 25} \\ \mathbf{0}_{10\times 3} & \mathbf{0}_{10\times 25} \\ \mathbf{0}_{25\times 3} & \boldsymbol{I}_{25} \end{bmatrix}$	$\begin{bmatrix} \boldsymbol{B}^* & \mathbf{0}_{15\times 25} \\ \mathbf{0}_{10\times 3} & \mathbf{0}_{10\times 25} \\ \mathbf{0}_{25\times 3} & -\tilde{\boldsymbol{B}}_{25\times 25} \end{bmatrix}$	$\begin{bmatrix} (1+\rho)\alpha_1 \tilde{\boldsymbol{\iota}} \\ \mathbf{0}_{45\times 1} \end{bmatrix}$

Notes: see notes in Table 3.1. Additionally, we define $b = 1$ ($b \sim U(0, 0.2)$) and $\tilde{\boldsymbol{B}}$ as a diagonal matrix with $b_{ii} = 1$ ($b_{ii} \sim U(0, 0.2)$) in the absence (presence) of persistence, and $\boldsymbol{B}^* = (\mathbf{1}_{3\times 3} \otimes \tilde{\boldsymbol{\iota}})$.

the ADL-ADF model. Based on the general DGP (3.23), we distinguish four different cases, corresponding to: (i) different orders of the dependent variable ($I(0)/I(1)$) and (ii) different degrees of persistence in the stationary variables (low/high). The choice to include varying degrees of persistence is motivated by the conjecture that the performance of the pre-testing procedure incorporated in the ADL-ADF model may deteriorate when the degree of persistence increases, which in turn translates to a decrease in the overall performance of the procedure.

The parameter settings for the varying DGPs, displayed in Table 3.2, are chosen such that they allow for a subset of stationary variables in the system. In particular, we first consider a scenario in which the dependent variable itself admits a stationary autoregressive representation in levels. In addition, based on their cross-sectional ordering, the first 15 variables after y are cointegrated based on three cointegrating vectors, the next 10 variables are non-cointegrated random walks, and the last 24 variables all admit a stationary autoregressive structure in levels. The degree of persistence in the stationary variables is regulated by the diagonal matrix $\tilde{\boldsymbol{B}}$ in $\boldsymbol{B}$, with elements $b_{ii} = 1$ in the low persistence case and $b_{ii} \sim U(0, 0.2)$ in the high persistence case. It can be seen from the last column in Table 3.2, that in line with the stationarity of the dependent variable, the first element in $\boldsymbol{\delta}$ will always be equal to -1, whereas an additional five-dimensional cointegrating vector enters the single-equation model for positive values of a. For the scenario in which the dependent variable is integrated of order one, the first 15 variables (including y) are all cointegrated based on three cointegrating vectors, the next 10 variables are non-cointegrated random walks, whereas the last 15 variables all admit a stationary autoregressive representation. The persistence in the stationary variables is regulated similar to the previous case. Now, however, it is clear from the last column in Table

3.2 that $\boldsymbol{\delta} \neq \mathbf{0}$ only if $a > 0$, such that lagged levels only enter the single-equation when y is cointegrated with its neighbouring variables. We display the performance of the models in Figure 3.2.

In the first two rows of Figure 3.2, corresponding to $y \sim I(0)$ and low persistence, SPECS correctly selects the lagged dependent variable in all simulation trials, such that the pseudo-power plot displays a constant line at 1. Interestingly, the PCS also seems constant around 35%. Upon closer inspection, we find that SPECS chooses an alternative representation of the single-equation model in which the contribution of the non-trivial cointegrating vector seems to be absorbed in the lagged level of the dependent variable. While the resulting model differs from the implied oracle model, which we indeed find to be accurately estimated by the OLS oracle procedure, the model choice seems to be motivated by a favourable bias-variance trade-off. In line with this conjecture, the nowcast performance of SPECS occasionally exceeds that of the oracle procedure in which a larger number of parameters must be estimated. Focussing on the ADL models, we observe that the standard ADL nowcasts are again inferior, whereas the ADL-ADF model seems to benefit from correct identification of the stationarity of the dependent variable, which is particularly relevant given that the dependent variable itself is a main component in the optimal forecast.[10] However, the nowcast accuracy of SPECS is almost identical to that of the ADL-ADF model, a finding that we interpret as reassuring and confirmatory of our claim that SPECS may be used without any pre-testing procedure. Moreover, the absence of strong persistence in the stationary variables idealizes the results of the ADL-ADF procedure. In typical macroeconomic applications many time series that are considered as I(0) display much slower mean reversion and, consequently, are more difficult to correctly identify as being stationary.[11] Accordingly, in the second row we display the result for a DGP where the stationary variables display more persistent behaviour. The performance of SPECS remains largely unaffected, whereas the nowcasting performance of the ADL-ADF model deteriorates drastically. We stress the relevance of this result, given that this estimation method in combination with a similar pre-testing procedure is fairly common practice. Somewhat surprisingly, the ADL model in differences nowcasts almost as well as SPECS for this particular setting. Overall, however, the nowcasts of SPECS remain the most accurate and, equally important, most stable across all specifications.

[10] The importance of correctly identifying the order of integration of the dependent variable returns in Chapter 5 as well.

[11] For example, the ten time series in the popular Fred-MD dataset which McCracken and Ng (2016) propose to be I(0), i.e. the series corresponding to a tcode of one, all display strong persistence or near unit root behaviour, with the smallest estimated AR(1) coefficient exceeding 0.86.

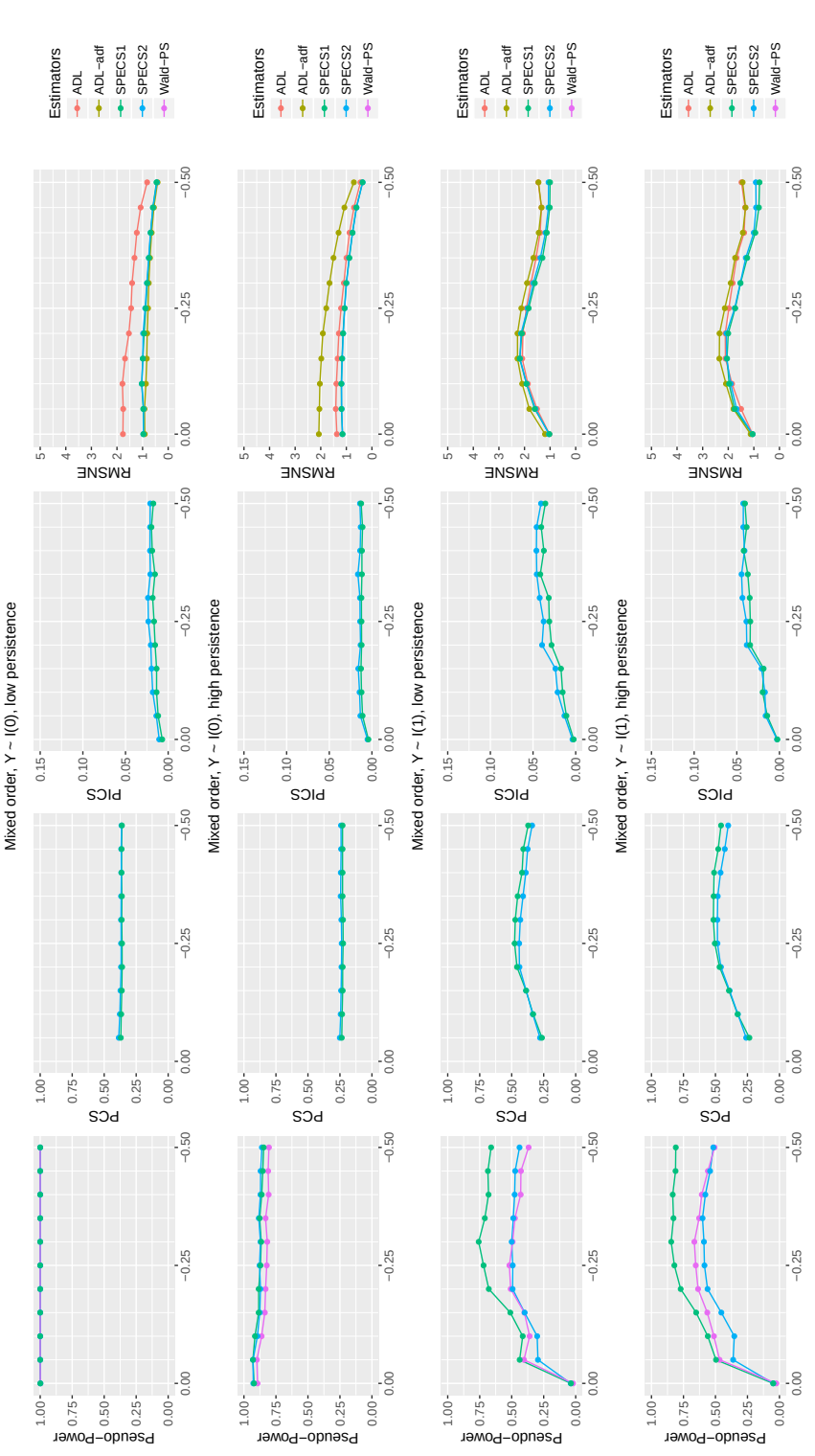

Figure 3.2: Pseudo-Power, Proportion of Correct Selection (PCS), Proportion of Incorrect Selection (PICS) and Root Mean Squared Nowcast Error (RMSNE) for four Mixed Order specifications. The adjustment rate multiplier α_1 is on the horizontal axis.

Table 3.3 Nowcasting performance on a DGP with a non-stationary factor.

	Root Mean Squared Nowcast Error		
	SPECS_1	SPECS_2	SPECS_1 - OLS
No Dynamics	1.07	1.11	0.99
Dynamics	1.02	1.02	1.01

This table reports the root mean squared nowcast errors relative to the ADL model.

Continuing the analysis of mixed order datasets, rows 3 and 4 of Figure 3.2 display the results for DGPs where the dependent variable is generated as being integrated of order one. The pseudo-power plot clearly reflects that $\boldsymbol{\delta} \neq \mathbf{0}$ only when $\alpha_1 > 0$. Furthermore, while SPECS performs well at removing the irrelevant variables, the relevant variables are not all selected correctly, resulting in somewhat lower values for the PCS metric. Nevertheless, the nowcast performance remains superior to that of the ADL model, especially in the presence of cointegration with fast adjustment rates.

3.4.3 A Dense Factor Model

Finally, to avoid idealizing the results through a choice of DGPs that suits our procedure, we consider a more adverse setting by generating the data with a non-stationary factor structure, while allowing for contemporaneous correlation and dynamic structures in both the error processes driving the 'observable' data and the idiosyncratic component in the factor structure. The DGP that we adopt corresponds to setting III in Palm et al. (2011, p. 92). For completeness, the DGP is given by

$$\boldsymbol{z}_t = \boldsymbol{\lambda} f_t + \boldsymbol{\omega}_t,$$

where $\boldsymbol{z}_t$ is a (50×1) time series process and f_t is a single scalar factor. Furthermore,

$$\begin{aligned} f_t &= \phi f_{t-1} + \zeta_t, \\ \omega_{i,t} &= \theta_i \omega_{i,t-1} + v_{i,t} \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{v}_t &= \boldsymbol{A}_1 \boldsymbol{v}_{t-1} + \boldsymbol{\epsilon}_{1,t} + \boldsymbol{B}_1 \boldsymbol{\epsilon}_{1,t-1}, \\ \zeta_t &= \alpha_2 \zeta_{t-1} + \epsilon_{2,t} + \beta_2 \epsilon_{2,t-1}, \end{aligned}$$

where $\boldsymbol{\epsilon}_{1,t} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and $\epsilon_{2,t} \sim \mathcal{N}(0, 1)$. The comparison focuses exclusively on

the nowcasting performance for a setting without dynamics ($\boldsymbol{A}_1 = \boldsymbol{B}_1 = \mathbf{0}$ and $\alpha_2 = \beta_2 = 0$) and a setting with dynamics ($\alpha_2 = \beta_2 = 0.4$). The construction of $\boldsymbol{A}_1$ and $\boldsymbol{B}_1$ is analogous to Palm et al. (2011, p. 93). We report the RMSNEs of SPECS relative to the ADL in Table 3.3. Given that the single-equation model is misspecified in this setup, it is unreasonable to expect SPECS to outperform. Indeed, we observe that the RMSNEs are all very close to one and, while in most cases the ADL model performs slightly better, the difference seems negligible. Hence, the risk of using SPECS to estimate a misspecified model in the sense considered here, does not seem to be higher than the use of the alternative ADL model, whereas the relative merits of SPECS when applied to a wide range of correctly specified models are evident from the first part of the simulations.

3.5 Empirical Application

Inspired by Choi and Varian (2012), we consider the possibility of nowcasting Dutch unemployment with our methods based on Google Trends data. Google Trends are hourly updated time series consisting of normalized indices depicting the volume of search queries entered in Google originating from a certain geographical area that were entered into Google. The Dutch unemployment rates are made available by Statistics Netherlands, an autonomous administrative body focussing on the collection and publication of statistical information. These rates are published on a monthly basis with new releases being made available on the 15th of each new month. This misalignment of publication dates clearly illustrate a practically relevant scenario where improvements upon forward looking predictions of Dutch unemployment rates may be obtained by utilizing contemporaneous Google Trends series.

We collect a novel dataset containing seasonally unadjusted Dutch unemployment rates from the website of Statistics Netherlands[12] and a set of manually selected Google Trends time series containing unemployment related search queries, such as 'Vacancy', 'Resume' and 'Unemployment Benefits'. The dataset comprises of monthly observations ranging from January 2004 to December 2017. While the full dataset contains 100 unique search queries, a number of these contain zeroes for large sub-periods indicating insufficient search volumes for those particular series. Consequently, we remove all series that are perfectly correlated over any sub-period consisting of 20% of the total sample.[13]

The benchmark model we consider is an ADL model fitted to the differenced

[12] `http://statline.cbs.nl/StatWeb/publication/?VW=T&DM=SLEN&PA=80479eng&LA=EN`

[13] The dataset is available with the *R* code at https://sites.google.com/view/etiennewijler.

Table 3.4 Number of parameters.

p	N'	ADL-ADF	SPECS_1	SPECS_2
1	262	1.27	0.99	1.07
3	436	1.06	0.82*	0.88
6	697	0.90	0.90	0.84*

This table reports the number of parameters estimated, N', as well as the Mean-Squared Nowcast Error relative to the ADL model for varying number of lagged differences p. We use * to denote rejection by the Diebold-Mariano test at the 10% significance level.

data. In detail, let y_t and $\boldsymbol{x}_t$ be the scalar unemployment rate and the vector of Google Trends series observed at time t, respectively, and define $\boldsymbol{z}_t = (y_t, \boldsymbol{x}_t')'$. The benchmark ADL estimator fits

$$\Delta y_t = \boldsymbol{\pi}_0' \Delta \boldsymbol{x}_t + \sum_{j=1}^{p} \boldsymbol{\pi}_j' \Delta \boldsymbol{z}_{t-j} + \epsilon_t.$$

However, this estimator ignores the order of integration of individual time series by differencing the whole dataset, while it is common practice to transform individual series to stationarity based on a preliminary test for unit roots. Hence, we include another ADL model where the decision to difference is based on a preliminary ADF test referred to as ADL-ADF.[14] Finally, SPECS estimates

$$\Delta y_t = \boldsymbol{\delta}' \boldsymbol{z}_{t-1} + \boldsymbol{\pi}_0' \Delta \boldsymbol{x}_t + \sum_{j=1}^{p} \boldsymbol{\pi}_j' \Delta \boldsymbol{z}_{t-j} + \epsilon_t.$$

All tuning parameters are obtained by time series cross-validation (Hyndman, 2016) and we use $k_\delta = 1.1$ which performed well based on a preliminary analysis.[15] The first nowcast is made by fitting the models on a window containing the first two-thirds of the complete sample, i.e. $t = 1, \ldots, T_c$ with $T_c = \lceil \frac{2}{3} T \rceil$, based on which the nowcast for Δy_{T_c+1} is produced. This procedure is repeated by rolling the window forward by one observation until the end of the sample is reached, producing a total of 54 pseudo out-of-sample nowcasts. In Table 3.4 we report the MSNE relative to the ADL model for $p = 1, 3, 6$.

[14] We note that none of the time series were found to be integrated of order 2. The outcome of the ADF test is reported for each time series in Appendix 3.B.2.

[15] We compared the nowcast accuracy for varying $k_\delta \in [0, 2]$ and observed that the lowest nowcast accuracy was obtained for $k_\delta = 1.1$, whereas for values of $k_\delta > 1.5$ almost all lagged levels were consistently excluded. In the latter case, the nowcast accuracy of SPECS was similar to that of the ADL benchmark.

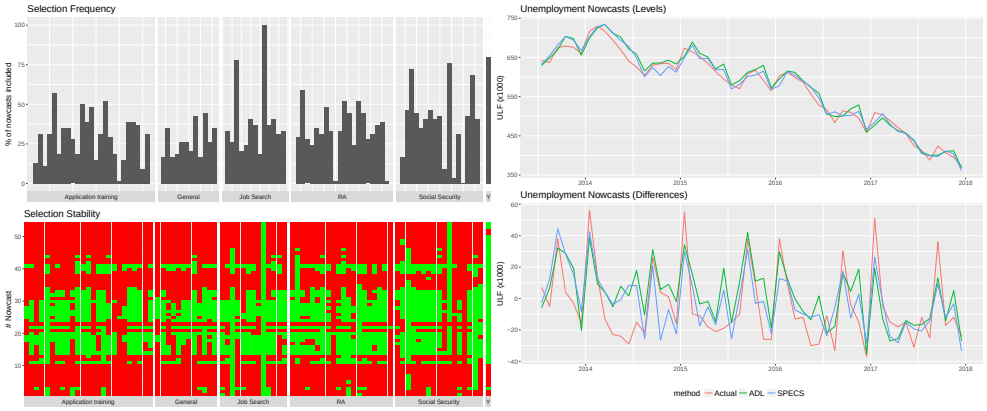

Figure 3.3: *Top-left*: Selection frequency, measured as the percentage of all nowcasts the variable was selected. *Bottom-left*: Selection stability with green indicating a variable was included in the nowcast model and red indicating exclusion. *Right*: Actual versus predicted unemployed labour force (ULF) in levels and differences.

The ADL-ADF estimator does not perform better than the regular ADL model for $p = 1, 3$, indicating that the potential for errors in pre-testing might lead to unfavourable results. SPECS performs well and is able to obtain smaller mean-squared nowcast errors than the ADL benchmark across almost all specifications, with the combination SPECS_2 and $p = 1$ being the exception. Moreover, for SPECS_1 ($p = 3$) and SPECS_2 ($p = 6$), we find the differences in MSNE to be significant at the 10% level according to the Diebold-Mariano test. The overall (unreported) MSNE is lowest for the SPECS_1 estimator based on $p = 3$ lagged differences. Given that the addition of lagged levels to the models improves the nowcast performance, the premise of cointegrating relationships between Dutch unemployment rates and Google Trends series seems likely. To further explore the presence of cointegration among our time series we group our variables in five categories; (1) Application, (2) General, (3) Job Search, (4) Recruitment Agencies (RA) and (5) Social Security. We narrow down our focus to the nowcasts of models with three lagged difference included, $p = 3$, estimated by SPECS_1. In Figure 3.3 we visually display the share of nowcasts in which the lagged levels of each variable are included in the estimated model. In addition, it depicts the selection stability of those variables, where a green colour indicates that a given variables is included in a given nowcast, and red vice versa. The figure also displays the actual unemployment rates compared to the nowcasted values.

Figure 3.3 highlights that only few variables are consistently selected for all nowcasts, although in each category we can distinguish some variables that are included at

higher frequencies. The variable whose lagged levels are always selected is 'Vakantiebaan', which is a search query for a temporary job during the summer holiday. We postulate that this variable is selected by SPECS to account for seasonality in the Dutch unemployment rates. In an unreported exercise we estimate the model with the addition of a set of eleven unpenalized dummies representing different months of the year. While in this experiment the variable 'Vakantiebaan' is never selected, the mean squared nowcast error increases substantially. Hence, we opt to adhere to our standard model under the caveat that for at least one of the lagged levels included, seasonality effects rather than cointegration seem a more appropriate explanation for its inclusion. Other frequently included variables are queries for vacancies ('uwv.vacatures', 78%), unemployment ('werkloos', 76%) and social benefits ('ww uitkering', 72%), where the stated percentages indicate the percentage of nowcast models in which the respective variables are selected. Furthermore, the last bar represents the frequency in which the lagged level of the Dutch unemployment rate is selected, which occurs for 43 out of 54 nowcasts (80%). The frequent selection of the lagged level of unemployment rates in conjunction with the other lagged levels is indicative of the presence of cointegration among unemployment and Google Trends series. However, we do not attach any structural meaning to the found equilibria based on the difficulty of interpretation when one does not assume the presence of weak exogeneity.

To gain insights into the temporal stability of our estimator, we visualize the selection stability in the bottom-left part of Figure 3.3. Generally, for the early and later period of the sample very few time series enter the model in levels, whereas for the middle part of the sample the majority of variables are selected. The exact reason for these patterns to occur is unknown and raises questions on the stability of Google trends as informative predictors of Dutch unemployment rates. Feasible explanations include structural instability in the DGP, seasonality effects or data idiosyncrasies. However, there are additional peculiarities specific to the use of Google trends such as normalization, data hubris and search algorithm dynamics, all of which might result in unstable performance (cf. Lazer et al., 2014). Since the focus of this application is not on a structural analysis of the relation between Google Trends and unemployment rates, we consider this issue outside the scope of the chapter. Instead, we focus on the relative empirical performance of our methods, which, notwithstanding the aforementioned caveats, we deem convincingly favourable for SPECS. Finally, on the right of Figure 3.3, we display the realized and predicted unemployment rates in levels and differences. Both the penalized ADL model and SPECS seem to follow the actual unemployment rates with reasonable accuracy, with the largest nowcast errors occurring in the first half of 2014. Prior to this period the unemployment rates had

been steadily rising in the aftermath of the economic recession, whereas 2014 marks the start of a recovery period. Given that the models are fit on historical data, it is natural that the estimators overestimate the unemployment rate shortly after the start of the economic recovery. Perhaps not entirely coincidental, the start of the period over which the majority of lagged levels are included by SPECS coincides with this recovery period as well, thereby hinting towards structural instability in the DGP as a plausible cause for the observed selection instability.

3.6 Conclusion

In this chapter we propose the use of SPECS as an automated approach to cointegration modelling. SPECS is an intuitive estimator that applies penalized regression to a conditional error-correction model. We show that SPECS possesses the oracle property and is able to consistently select the long-run and short-run dynamics in the underlying DGP. A simulation exercise confirms strong selective and predictive capabilities in both low and high dimensions with impressive gains over a benchmark penalized ADL model that ignores cointegration in the dataset. The assumption of weak exogeneity is important for efficient estimation and interpretation of the model. However, while our estimator is not entirely insensitive to this assumption, the simulation results demonstrate that the selective capabilities remain adequate and the nowcasting performance remains superior to the benchmark. Finally, we consider an empirical application in which we nowcast the Dutch unemployment rate with the use of Google Trends series. Across all three different dynamic specifications considered, SPECS attains higher nowcast accuracy, thus confirming the results in our simulation study. As a result, we believe that our proposed estimator, which is easily implemented with readily available tools at low computational cost, offers a valuable tool for practitioners by enabling automated model estimation on relatively large and potentially non-stationary datasets and, most importantly, allowing to take into account potential (co)integration without requiring pre-testing procedures.

The use of a fixed-dimensional asymptotic framework can be considered as a limitation that applies to this chapter. While the fixed-dimensional framework allows the theoretical results to be derived under fewer and more intuitive assumptions, it is not informative of the behaviour one may expect in high-dimensional applications. However, the simulation exercise and the empirical application provide promising results which seem to indicate that the theoretical properties of SPECS carry over to a high-dimensional asymptotic framework. We consider this issue in the following chapter.

Appendix 3.A Proofs

3.A.1 Preliminary Results

Similar to (3.8), we write the conditional error correction model in matrix notation as

$$\Delta \boldsymbol{y} = \boldsymbol{Z}_{-1}\boldsymbol{\delta} + \boldsymbol{W}\boldsymbol{\pi} + \boldsymbol{\iota}\mu_0 + \bar{\boldsymbol{t}}\tau_0 + \boldsymbol{\epsilon}_y,$$

where by construction $\mathbb{E}(\boldsymbol{\epsilon}_t \epsilon_{y,t}) = \boldsymbol{0}$. Following the discussion in Section 3.3.2, we may equivalenty write this as

$$\Delta \boldsymbol{y} = \boldsymbol{Z}_{1,-1}\boldsymbol{\delta}_1 + \boldsymbol{W}_2\boldsymbol{\pi}_2 + \boldsymbol{\iota}\mu_0 + \bar{\boldsymbol{t}}\tau_0 + \boldsymbol{\epsilon}_y,$$

where $\boldsymbol{Z}_{1,-1}$ contains the subset of variables in $\boldsymbol{Z}_{-1}$ that are $I(1)$ and $\boldsymbol{W}_2 = (\boldsymbol{Z}_{2,-1}, \boldsymbol{W})$ with $\boldsymbol{Z}_{2,-1}$ the subset of $I(0)$ variables. For notational convencience we proceed under the assumption that all variables are integrated of order one such that $\boldsymbol{Z}_{-1} = \boldsymbol{Z}_{1,-1}$. We stress, however, that this assumption is without loss of generality, as one may replace the matrices in the proof below by their decomposed variants without additional complications. Under Assumption 3.2, the moving average representation of the N-dimensional time series $\boldsymbol{z}_t$ is given by

$$\boldsymbol{Z}_{-1} = \boldsymbol{S}_{-1}\boldsymbol{C}' + \boldsymbol{\iota}\boldsymbol{\mu}' + \bar{\boldsymbol{t}}\boldsymbol{\tau}' + \boldsymbol{U}_{-1}, \tag{3.A.1}$$

where $\boldsymbol{S}_{-1} = (\boldsymbol{s}_0, \dots, \boldsymbol{s}_{T-1})'$, with $\boldsymbol{s}_t = \sum_{i=1}^{t} \boldsymbol{\epsilon}_i$,

$$\boldsymbol{C} = \boldsymbol{B}_\perp \left(\boldsymbol{A}_\perp' \left(\boldsymbol{I}_N - \sum_{j=1}^{p} \boldsymbol{\Phi}_j \right) \boldsymbol{B}_\perp \right)^{-1} \boldsymbol{A}_\perp,$$

and $\boldsymbol{U}_{-1} = (\boldsymbol{u}_0, \dots, \boldsymbol{u}_{T-1})'$, with $\boldsymbol{u}_t = \boldsymbol{C}(L)\boldsymbol{\epsilon}_t + \boldsymbol{z}_0$ consisting of a linear process plus initial conditions.

We first present a number of useful intermediary results that will aid the proofs of our main results. The first of such results details the weak convergence of integrated processes. Based on Assumption 3.1, the following results are well-known in the literature.

Lemma 3.A.1. *Let $\boldsymbol{B}(r)$ denote a Brownian Motion with covariance matrix $\boldsymbol{\Sigma}$ and define $\boldsymbol{D} = (\boldsymbol{\iota}, \bar{\boldsymbol{t}})$ and $\boldsymbol{M}_D = \boldsymbol{I}_T - \boldsymbol{D}(\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'$. Then, under Assumption 3.1,*

(a) $T^{-2}\boldsymbol{S}_{-1}'\boldsymbol{S}_{-1} \xrightarrow{d} \int_0^1 \boldsymbol{B}(r)\boldsymbol{B}(r)'dr$

(b) $T^{-3/2}\boldsymbol{S}'_{-1}\boldsymbol{\iota} \xrightarrow{d} \int_0^1 \boldsymbol{B}(r)dr$

(c) $T^{-5/2}\boldsymbol{S}'_{-1}\bar{\boldsymbol{t}} \xrightarrow{d} \int_0^1 r\boldsymbol{B}(r)dr$

(d) $T^{-1}\boldsymbol{S}'_{-1}\boldsymbol{\epsilon}_y \xrightarrow{d} \int_0^1 \boldsymbol{B}(r)dB_{\epsilon_y}(r)$

(e) $T^{-3/2}\boldsymbol{S}'_{-1}\boldsymbol{U}_{-1} \xrightarrow{d} \left(\int_0^1 \boldsymbol{B}(r)dr\right)\boldsymbol{z}'_0$

(f) $T^{-1}\boldsymbol{U}'_{-1}\boldsymbol{U}_{-1} \xrightarrow{p} \sum_{j=0}^{\infty}\boldsymbol{C}_j\boldsymbol{\Sigma}\boldsymbol{C}'_j$.

In addition, these results carry through for $\boldsymbol{S}^*_{-1} = \boldsymbol{M}_D\boldsymbol{S}_{-1}$ *by replacing* $\boldsymbol{B}(r)$ *for* $\boldsymbol{B}^*(r) = \boldsymbol{B}(r) - \int_0^1 \boldsymbol{B}(s)ds - 12\left(r-\frac{1}{2}\right)\int_0^1\left(s-\frac{1}{2}\right)\boldsymbol{B}(s)ds$ *in the corresponding limit distributions.*

Proof. Under Assumption 3.1, Phillips and Solo (1992) show that $\boldsymbol{\epsilon}_t$ satisfies a multivariate invariance principle. Consequently, the convergence results (a)-(e) are directly implied by Lemma 2.1 in Park and Phillips (1989), whereas (f) is a standard result for linear processes (e.g. Brockwell and Davis, 1991, p. 404). The claim that the convergence holds true after de-meaning and de-trending, i.e. after pre-multiplication of the data matrix by $\boldsymbol{M}_D$, can be found in most standard time series textbooks, see for example Davidson (2000, p. 354). ■

Absent of cointegration in the data, the matrix $\boldsymbol{C}$ will be of full rank. In this setting, the following convergence results are well-established in the literature.

Lemma 3.A.2. *Let* $\boldsymbol{M}_D$ *be defined as in Lemma 3.A.1. Then, under Assumptions 3.1 and 3.2,*

(a) $T^{-2}\boldsymbol{Z}'_{-1}\boldsymbol{M}_D\boldsymbol{Z}_{-1} \xrightarrow{d} \boldsymbol{C}\left(\int_0^1 \boldsymbol{B}^*(r)\boldsymbol{B}^{*\prime}(r)dr\right)\boldsymbol{C}'$,

(b) $T^{-3/2}\boldsymbol{Z}'_{-1}\boldsymbol{M}_D\boldsymbol{W} \xrightarrow{p} 0$,

(c) $T^{-1}\boldsymbol{W}'\boldsymbol{M}_D\boldsymbol{W} \xrightarrow{p} \boldsymbol{\Sigma}_w$,

(d) $T^{-1}\boldsymbol{Z}'_{-1}\boldsymbol{M}_D\boldsymbol{\epsilon}_y \xrightarrow{d} \int_0^1 \boldsymbol{B}^*(r)dB_{\epsilon_y}(r)$,

(e) $T^{-1/2}\boldsymbol{W}'\boldsymbol{M}_D\boldsymbol{\epsilon}_y \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\epsilon_y}\boldsymbol{\Sigma}_w\right)$,

where $\boldsymbol{B}^*(r)$ *as in Lemma 3.A.1.*

Proof. These results are standard and details of the proof are omitted. Briefly, one can plug in the definitions of the matrices $\boldsymbol{Z}_{-1}$ and $\boldsymbol{W}$ based on (3.A.1), and apply Lemma 3.A.1 to show the results (a)-(d). Result (e) follows from an application of a central limit theorem for linear process as in Theorem 3.4 in Phillips and Solo (1992). ■

When cointegration is present in the data, the matrix $\boldsymbol{C}$ will be of rank $N - r$, which will be problematic in applications where its inverse is required. A workaround is to transform the system into a stationary and non-stationary component. From (3.A.1), it follows that

$$\boldsymbol{Z}_{-1}\boldsymbol{B} = \boldsymbol{\iota}\boldsymbol{\mu}'\boldsymbol{B} + \bar{\boldsymbol{t}}\boldsymbol{\tau}'\boldsymbol{B} + \boldsymbol{U}_{-1}\boldsymbol{B}$$

is a (trend-)stationary process and

$$\boldsymbol{Z}_{-1}\boldsymbol{A}_\perp = \boldsymbol{S}^{-1}\boldsymbol{C}'\boldsymbol{A}_\perp + \boldsymbol{\iota}\boldsymbol{\mu}'\boldsymbol{A}_\perp + \bar{\boldsymbol{t}}\boldsymbol{\tau}'\boldsymbol{A}_\perp + \boldsymbol{U}_{-1}\boldsymbol{A}_\perp$$

contains the stochastic trends.[16] Accordingly, define the linear transformation

$$\boldsymbol{Q} := \begin{bmatrix} \boldsymbol{B}' & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_M \\ \boldsymbol{A}'_\perp & \boldsymbol{0} \end{bmatrix} \text{ with } \boldsymbol{Q}^{-1} = \begin{bmatrix} \boldsymbol{A}(\boldsymbol{B}'\boldsymbol{A})^{-1} & \boldsymbol{0} & \boldsymbol{B}_\perp(\boldsymbol{A}'_\perp\boldsymbol{B}_\perp)^{-1} \\ \boldsymbol{0} & \boldsymbol{I}_M & \boldsymbol{0} \end{bmatrix},$$

and let $\boldsymbol{V} = (\boldsymbol{Z}_{-1}, \boldsymbol{W})$. Then,

$$\boldsymbol{V}\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Z}_{-1}\boldsymbol{B} & \boldsymbol{W} & \boldsymbol{Z}_{-1}\boldsymbol{A}_\perp \end{bmatrix} = \begin{bmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_2 \end{bmatrix},$$

with $\boldsymbol{V}_1 = (\boldsymbol{Z}_{-1}\boldsymbol{B}, \boldsymbol{W})$. We maintain the convention that for the case $r = N$, we define $\boldsymbol{B}_\perp = \boldsymbol{A}_\perp = \boldsymbol{0}$ and $\boldsymbol{V} = \boldsymbol{V}_1$. Based on this decomposition, we recall a number of convergence results under the remark that the results involving $\boldsymbol{V}_2$ are relevant only for the case $r < N$.

Lemma 3.A.3. *Let $\boldsymbol{M}_D$ be defined as in Lemma 3.A.1. Then, under Assumptions 3.1 and 3.2,*

(a) $T^{-2}\boldsymbol{V}_2'\boldsymbol{M}_D\boldsymbol{V}_2 \xrightarrow{d} \boldsymbol{A}'_\perp\boldsymbol{C}\left(\int_0^1 \boldsymbol{B}^*(r)\boldsymbol{B}^{*\prime}(r)dr\right)\boldsymbol{C}'\boldsymbol{A}_\perp$

(b) $T^{-3/2}\boldsymbol{V}_2'\boldsymbol{M}_D\boldsymbol{V}_1 \xrightarrow{p} \boldsymbol{0}$

(c) $T^{-1}\boldsymbol{V}_1'\boldsymbol{M}_D\boldsymbol{V}_1 \xrightarrow{p} \boldsymbol{\Sigma}_{V_1}$

(d) $T^{-1}\boldsymbol{V}_2'\boldsymbol{M}_D\boldsymbol{\epsilon}_y \xrightarrow{d} \boldsymbol{A}'_\perp\boldsymbol{C}\left(\int_0^1 \boldsymbol{B}^*(r)dB_{\epsilon_y}(r)\right)$

(e) $T^{-1/2}\boldsymbol{V}_1'\boldsymbol{M}_D\boldsymbol{\epsilon}_y \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\epsilon_y}\boldsymbol{\Sigma}_{V_1}\right)$

Proof. These results correspond to Lemma 1 in Ahn and Reinsel (1990) and we refer the reader to the original paper for their proofs. ■

[16] Note that $\boldsymbol{C}'\boldsymbol{A}_\perp$ simplifies to $\boldsymbol{A}_\perp$ when $\boldsymbol{\Phi}_j = 0$ for $j = 1, \ldots, p$.

The final preliminary result that will be used is an extension of the Frisch-Wraugh-Lovell theorem to penalized regression.

Lemma 3.A.4. *Let $\boldsymbol{M}_D$ be defined as in Lemma 3.A.1 and consider the solutions to the following two lasso regressions:*

$$\left(\hat{\boldsymbol{\gamma}}', \hat{\boldsymbol{\theta}}'\right)' = \underset{\boldsymbol{\gamma},\boldsymbol{\theta}}{arg\ min} \left\|\Delta \boldsymbol{y} - \boldsymbol{V}\boldsymbol{\gamma} - \boldsymbol{D}\boldsymbol{\theta}\right\|_2^2 + P_\lambda(\boldsymbol{\gamma}), \tag{3.A.2}$$

$$\breve{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{arg\ min} \left\|\boldsymbol{M}_D \Delta y - \boldsymbol{M}_D \boldsymbol{V}\boldsymbol{\gamma}\right\|_2^2 + P_\lambda(\boldsymbol{\gamma}), \tag{3.A.3}$$

where

$$P_\lambda(\boldsymbol{\gamma}) = \lambda_G \left(\sum_{i=1}^{N} |\gamma_i|^2\right)^{1/2} + \sum_{i=1}^{N} \lambda_{2,i} |\gamma_i| + \sum_{j=1}^{M} \lambda_{3,j} |\gamma_{N+j}| .$$

Based on (3.A.2) *and* (3.A.3) *we have*

(i) $\hat{\boldsymbol{\gamma}} = \breve{\boldsymbol{\gamma}}$;

(ii) $\hat{\boldsymbol{\theta}} = (\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'(\Delta \boldsymbol{y} - \boldsymbol{V}'\hat{\boldsymbol{\gamma}})$.

Proof of Lemma 3.A.4. The proof is provided in Yamada (2017) for the standard lasso. In our case the only difference is the addition of the derivative of the group penalty in the subgradient vector. Once this contribution is added the proof is entirely analogous. ■

3.A.2 Proofs of Theorems

Proof of Theorem 3.1. The proof largely follows along the lines of Liao and Phillips (2015). Recall from (3.9) that we obtain the standardized estimates $\hat{\gamma}^s$ by minimizing

$$G_T\left(\boldsymbol{\gamma}^s, \boldsymbol{\theta}\right) = \left\|\Delta \boldsymbol{y} - \tilde{\boldsymbol{V}}\boldsymbol{\gamma}^s - \boldsymbol{D}\boldsymbol{\theta}\right\|_2^2 + P_\lambda(\boldsymbol{\gamma}^s),$$

which by Lemma 3.A.4 are equivalent to those obtained from minimizing

$$G_T(\boldsymbol{\gamma}^s) = \left\|\boldsymbol{M}_D\left(\Delta \boldsymbol{y} - \tilde{\boldsymbol{V}}\boldsymbol{\gamma}^s\right)\right\|_2^2 + P_\lambda(\boldsymbol{\gamma}^s), \tag{3.A.4}$$

where we defined $\tilde{\boldsymbol{V}} = \boldsymbol{V}\boldsymbol{\Sigma}_V^{-1}$ and $\boldsymbol{\gamma}^s = \boldsymbol{\Sigma}_V\boldsymbol{\gamma}$, with $\boldsymbol{\Sigma}_V = \text{diag}(\boldsymbol{\Sigma}_Z, \boldsymbol{\Sigma}_W)$ a diagonal weighting matrix, which results in the decomposition $\boldsymbol{\gamma}^s = (\boldsymbol{\delta}^{s\prime}, \boldsymbol{\pi}^{s\prime})' = (\boldsymbol{\delta}'\boldsymbol{\Sigma}_Z, \boldsymbol{\pi}'\boldsymbol{\Sigma}_W)'$. By construction we have $G_T(\hat{\boldsymbol{\gamma}}^s) < G_T(\boldsymbol{\gamma}^s)$, from which it follows

that

$$(\hat{\boldsymbol{\gamma}}^s - \boldsymbol{\gamma}^s)'\tilde{\boldsymbol{V}}'\boldsymbol{M}_D\tilde{\boldsymbol{V}}(\hat{\boldsymbol{\gamma}}^s - \boldsymbol{\gamma}^s) - 2(\hat{\boldsymbol{\gamma}}^s - \boldsymbol{\gamma}^s)'\tilde{\boldsymbol{V}}'\boldsymbol{M}_D\boldsymbol{\epsilon}_y \leq P_\lambda(\boldsymbol{\gamma}^s) - P_\lambda(\hat{\boldsymbol{\gamma}}^s),$$

which is equivalent to

$$(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})'\boldsymbol{V}'\boldsymbol{M}_D\boldsymbol{V}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) - 2(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})'\boldsymbol{V}'\boldsymbol{M}_D\boldsymbol{\epsilon}_y \leq P_\lambda(\boldsymbol{\gamma}^s) - P_\lambda(\hat{\boldsymbol{\gamma}}^s). \tag{3.A.5}$$

The strategy to derive consistency of the estimators consists of appropriately bounding both sides of (3.A.5) from which the results in Theorem 3.1 can be obtained. We first proceed under the assumption that there is no cointegration present in the underlying DGP, i.e. $\boldsymbol{\delta} = \boldsymbol{0}$. Define the scaling matrix $\boldsymbol{D}_T = \text{diag}(T\boldsymbol{I}_N, \sqrt{T}\boldsymbol{I}_M)$. Then, a lower bound for the first left-hand side term of (3.A.5) is given by

$$(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\boldsymbol{D}_T\boldsymbol{D}_T^{-1}\boldsymbol{V}'\boldsymbol{M}_D\boldsymbol{V}\boldsymbol{D}_T^{-1}\boldsymbol{D}_T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \geq \|\boldsymbol{D}_T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\|_2^2\,\phi_{\min},$$

where $\phi_{\min}$ is the smallest eigenvalue of $\boldsymbol{D}_T^{-1}\boldsymbol{V}'\boldsymbol{M}_D\boldsymbol{V}\boldsymbol{D}_T^{-1}$. Let $\boldsymbol{A}$ be a $(N \times N)$ matrix and define $\rho_{\min}(\boldsymbol{A}) : \mathbb{R}^{N\times N} \to \mathbb{C}$ as the function that extracts its minimum eigenvalue. Then, by the continuous mapping theorem, it follows that

$$\phi_{\min} \xrightarrow{d} \rho_{\min}\left(\begin{bmatrix} \boldsymbol{C}\left(\int_0^1 \boldsymbol{B}^*(r)\boldsymbol{B}^{*\prime}(r)dr\right)\boldsymbol{C}' & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_W \end{bmatrix}\right) > 0, \quad \text{a.s.} \tag{3.A.6}$$

The almost sure positiveness of the minimum eigenvalue is motivated as follows. Absent of cointegration, $\boldsymbol{C}$ is full rank and $\int_0^1 \boldsymbol{B}^*(r)\boldsymbol{B}^{*\prime}(r)dr \succ 0$ almost surely by Lemma A2 in Phillips and Hansen (1990), such that $\boldsymbol{C}\left(\int_0^1 \boldsymbol{B}^*(r)\boldsymbol{B}^{*\prime}(r)dr\right)\boldsymbol{C}' \succ 0$. Additionally, $\boldsymbol{\Sigma}_W \succ 0$ as a consequence of Assumption 3.1. Then, as a direct consequence of (3.A.6), it also holds that $\mathbb{P}(\phi_{\min} > 0) \to 1$.

The second term in (3.A.5) is bounded by

$$(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})'\boldsymbol{D}_T\boldsymbol{D}_T^{-1}\boldsymbol{V}'\boldsymbol{M}_D\boldsymbol{\epsilon}_y \leq \|\boldsymbol{D}_T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\|_2 \left\|\boldsymbol{D}_T^{-1}\boldsymbol{V}'\boldsymbol{M}_D\boldsymbol{\epsilon}_y\right\|_2 = \|\boldsymbol{D}_T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\|_2\, a_T,$$

where $a_T = \left\|\boldsymbol{D}_T^{-1}\boldsymbol{V}'\boldsymbol{M}_D\boldsymbol{\epsilon}_y\right\|_2 = O_p(1)$ by Lemma 3.A.2.

Next, we derive an upper bound for the right-hand side of (3.A.5). For ease of exposition, we write $\lambda_{2,i} = \omega_{\delta,i}^{k_\delta}\lambda_{\delta,T}$ and $\lambda_{3,j} = \omega_{\pi,j}^{k_\pi}\lambda_{\pi,T}$. First, note that

$$\begin{aligned}\lambda_{G,T}\left(\|\boldsymbol{\delta}^s\|_2 - \left\|\hat{\boldsymbol{\delta}}^s\right\|_2\right) &\leq \lambda_{G,T}\left\|\hat{\boldsymbol{\delta}}^s - \boldsymbol{\delta}^s\right\|_2 \leq \lambda_{G,T}\|\hat{\boldsymbol{\gamma}}^s - \boldsymbol{\gamma}^s\|_2 \\ &\leq T^{-1/2}\lambda_{G,T}\|\boldsymbol{D}_T(\boldsymbol{\gamma}^s - \hat{\boldsymbol{\gamma}}^s)\|_2,\end{aligned}$$

where $T^{-1/2}\lambda_{G,T} \to 0$ by assumption. To bound the difference between the individual penalties, we define $\boldsymbol{\lambda}_\gamma = (\boldsymbol{\lambda}_2', \boldsymbol{\lambda}_3')'$ and $\boldsymbol{\lambda}_{S_\gamma}$ as an $(N+M)$-dimensional vector with $\lambda_{S_\gamma,i} = \lambda_{\gamma,i}\mathbb{1}\{\gamma_i \neq 0\}$. Then,

$$\begin{aligned}
&\sum_{i=1}^{N} \lambda_{2,i}\left(|\delta_i^s| - \left|\hat{\delta}_i^s\right|\right) + \sum_{j=1}^{M} \lambda_{3,j}\left(|\pi_j^s| - |\hat{\pi}_j^s|\right) \\
&\quad \leq \sum_{i\in S_\delta} \lambda_{2,i}\left(|\delta_i^s| - \left|\hat{\delta}_i^s\right|\right) + \sum_{j\in S_\pi} \lambda_{3,j}\left(|\pi_j^s| - |\hat{\pi}_j^s|\right) \\
&\leq \sum_{i\in S_\delta} \lambda_{2,i}\left|\hat{\delta}_i^s - \delta_i^s\right| + \sum_{j\in S_\pi} \lambda_{3,j}\left|\hat{\pi}_j^s - \pi_j^s\right| = \boldsymbol{\lambda}_{S_\gamma}'\boldsymbol{\Sigma}_V \boldsymbol{D}_T^{-1}\boldsymbol{D}_T\left|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\right| \\
&\quad \leq \left\|\boldsymbol{D}_T^{-1}\boldsymbol{\Sigma}_V\boldsymbol{\lambda}_{S_\gamma}\right\|_2 \left\|\boldsymbol{D}_T(\hat{\boldsymbol{\gamma}}^s - \boldsymbol{\gamma}^s)\right\|_2 .
\end{aligned}$$

Furthermore, it is straightforward to see that $\left\|\boldsymbol{D}_T^{-1}\boldsymbol{\Sigma}_V\boldsymbol{\lambda}_{S_\gamma}\right\|_2 = o_p(1)$ if

$$\frac{\lambda_{3,j}\sigma_{W,jj}}{\sqrt{T}} = \frac{\lambda_{\pi,T}\sigma_{W,jj}}{\sqrt{T}\left|\hat{\pi}_{OLS,j}^{k_\pi}\right|} = o_p(1),$$

for all $j \in S_\pi$. Since $\hat{\pi}_{OLS,j} \xrightarrow{p} \pi_j$ by the consistency of the OLS estimator, we require the condition $\frac{\lambda_{\pi,T}\sigma_{W,\max}}{\sqrt{T}} \xrightarrow{p} 0$.

Combining the bounds obtained thus far we can rewrite (3.A.5) as

$$\begin{aligned}
&\phi_{\min}\left\|\boldsymbol{D}_T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right\|_2^2 - 2a_T\left\|\boldsymbol{D}_T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right\|_2 \\
&\quad \leq \left(T^{-1/2}\lambda_G + \left\|\boldsymbol{D}_T^{-1}\boldsymbol{\Sigma}_V\boldsymbol{\lambda}_{S_\gamma}\right\|_2\right)\left\|\boldsymbol{D}_T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right\|_2 ,
\end{aligned}$$

from which it follows that

$$\left\|\boldsymbol{D}_T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right\|_2 \leq \phi_{\min}^{-1}2a_T + \phi_{\min}^{-1}\left(T^{-1/2}\lambda_G + \left\|\boldsymbol{D}_T^{-1}\boldsymbol{\Sigma}_V\boldsymbol{\lambda}_{S_\gamma}\right\|_2\right) = O_p(1),$$

which demonstrates the consistency of our estimator absent of cointegration.

Next, we assume there exists cointegration between the variables in the DGP, i.e. $\boldsymbol{\delta} \neq \mathbf{0}$. Let $\boldsymbol{Q}$ be defined as in (3.13) and define the scaling matrix $\boldsymbol{S}_T = \operatorname{diag}(\sqrt{T}\boldsymbol{I}_{M+r}, T\boldsymbol{I}_{N-r})$. By arguments analogous to the case without cointegration, we obtain a lower bound for the first left-hand side term of (3.A.5) as

$$(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})'\boldsymbol{Q}^{-1}\boldsymbol{S}_T\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}'\boldsymbol{M}_D\boldsymbol{V}\boldsymbol{Q}'\boldsymbol{S}_T^{-1}\boldsymbol{S}_T\boldsymbol{Q}'^{-1}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \leq \psi_{\min}\left\|\boldsymbol{S}_T\boldsymbol{Q}'^{-1}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right\|_2^2 ,$$

where $\psi_{\min}$ is the smallest eigenvalue of $\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}'\boldsymbol{M}_D\boldsymbol{V}\boldsymbol{Q}'\boldsymbol{S}_T^{-1}$. By Lemma 3.A.3 and

the continuous mapping theorem, we have

$$\psi_{\min} \xrightarrow{d} \rho_{\min}\left(\begin{bmatrix} \boldsymbol{\Sigma}_{V_1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A}_{\perp}'\boldsymbol{C}\left(\int_0^1 \boldsymbol{B}^*(r)\boldsymbol{B}^{*\prime}(r)dr\right)\boldsymbol{C}'\boldsymbol{A}_{\perp} \end{bmatrix}\right) > 0, \quad \text{a.s.}$$

The almost sure positiveness is implied by the fact that the matrix $\boldsymbol{\Sigma}_{V_1} \succ 0$ as a consequence of Assumption 3.1. Additionally, by Assumption 3.2, $\boldsymbol{A}_{\perp}'\boldsymbol{C}$ is an $(r \times N)$-dimensional matrix of full-row rank r and $\int_0^1 \boldsymbol{B}^*(r)\boldsymbol{B}^{*\prime}(r)dr \succ 0$ by Lemma A2 in Phillips and Hansen (1990). Consequently, $\mathbb{P}(\psi_{\min} > 0) \to 1$.

The second term of (3.A.5) is bounded by

$$\begin{aligned}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})'\boldsymbol{Q}^{-1}\boldsymbol{S}_T\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}'\boldsymbol{M}_D\boldsymbol{\epsilon}_y &\leq \left\|\boldsymbol{S}_T\boldsymbol{Q}'^{-1}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right\|_2 \left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}'\boldsymbol{M}_D\boldsymbol{\epsilon}_y\right\|_2 \\ &= \left\|\boldsymbol{S}_T\boldsymbol{Q}'^{-1}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right\|_2 b_T,\end{aligned}$$

where $b_T = O_p(1)$ according to Lemma 3.A.3. The bounds for the right-hand side of (3.A.5) are the same as for the case $\boldsymbol{\delta} = \mathbf{0}$, but with $\boldsymbol{D}_T$ replaced by $\boldsymbol{S}_T\boldsymbol{Q}'^{-1}$. In particular, we obtain

$$\lambda_{G,T}\left(\|\boldsymbol{\delta}^s\|_2 - \left\|\hat{\boldsymbol{\delta}}^s\right\|_2\right) \leq T^{-1/2}\lambda_{G,T}\left\|\boldsymbol{S}_T\boldsymbol{Q}'^{-1}(\boldsymbol{\gamma}^s - \hat{\boldsymbol{\gamma}}^s)\right\|_2,$$

and

$$\sum_{i=1}^{N}\lambda_{2,i}\left(|\delta_i^s| - \left|\hat{\delta}_i^s\right|\right) + \sum_{j=1}^{M}\lambda_{3,j}\left(|\pi_j^s| - |\hat{\pi}_j^s|\right) \leq \left\|S_T^{-1}Q\sigma_V\lambda_{S_\gamma}\right\|_2 \left\|S_TQ'^{-1}(\hat{\gamma}_i^s - \gamma_i^s)\right\|_2.$$

Furthermore, we can bound

$$\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}_V\boldsymbol{\lambda}_{S_\gamma}\right\|_2 \leq T^{-1/2}\left\|\boldsymbol{\lambda}_{S_\gamma}\right\|_2 \|\boldsymbol{\Sigma}_V\|_2 \|\boldsymbol{Q}\|_2,$$

which is easily seen to be bounded in probability when $\frac{\lambda_{3,j}\sigma_{W,jj}}{\sqrt{T}} = o_p(1)$, for $j \in S_\pi$, and

$$\frac{\lambda_{2,i}\sigma_{Z,ii}}{\sqrt{T}} = \frac{\lambda_{\delta,T}\sigma_{Z,ii}}{\sqrt{T}\left|\hat{\delta}_{OLS,i}^{k_\delta}\right|} = o_p(1),$$

for $i \in S_\delta$. Since $\hat{\delta}_{OLS,i} \xrightarrow{p} \delta_i$ by the consistency of the OLS estimator, we require the additional condition $\frac{\lambda_{\delta,T}\sigma_{Z,\max}}{\sqrt{T}} \xrightarrow{p} 0$.

Combining the bounds for the case $\boldsymbol{\delta} \neq \mathbf{0}$ we can rewrite (3.A.5) as

$$\begin{aligned}&\psi_{\min} \left\| \boldsymbol{S}_T \boldsymbol{Q}'^{-1} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \right\|_2^2 - 2 b_T \left\| \boldsymbol{S}_T \boldsymbol{Q}'^{-1} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \right\|_2 \\ &\leq \left(T^{-1/2} \lambda_G + \left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{\Sigma}_V \boldsymbol{\lambda}_{S_\gamma} \right\|_2 \right) \left\| \boldsymbol{S}_T \boldsymbol{Q}'^{-1} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \right\|_2 ,\end{aligned}$$

which can be rewritten as

$$\left\| \boldsymbol{S}_T \boldsymbol{Q}'^{-1} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \right\|_2 \leq \psi_{\min}^{-1} 2 b_T + \psi_{\min}^{-1} \left(T^{-1/2} \lambda_G + \left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{\Sigma}_V \boldsymbol{\lambda}_{S_\gamma} \right\|_2 \right) = O_p(1),$$

thereby completing the proof for the case of cointegration. ■

Proof of Theorem 3.2. We first proceed by deriving the selection consistency for the case $\boldsymbol{\delta} = \mathbf{0}$. Assume that $\hat{\delta}_i^s \neq 0$ is a minimizer of (3.9) and thus, by application of Lemma 3.A.4, also minimizes (3.A.4). Let $\boldsymbol{z}_i$ denote the i-th column vector of Z_{-1}. The first order conditions for $\hat{\delta}_i^s$ to be a minimum state

$$\frac{\partial G_T(\boldsymbol{\gamma}^s)}{\partial \delta_i^s}\Big|_{\boldsymbol{\gamma}^s = \hat{\boldsymbol{\gamma}}^s} = \tilde{\boldsymbol{z}}_i' \boldsymbol{M}_D \left(\Delta \boldsymbol{y} - \tilde{\boldsymbol{V}} \hat{\boldsymbol{\gamma}}^s \right) - \frac{\lambda_G}{2} \hat{\delta}_i^s \left\| \hat{\boldsymbol{\delta}}^s \right\|_2^{-1} - \frac{\lambda_{2,i} \operatorname{sign}(\hat{\delta}_i^s)}{2} = 0.$$

After multiplying by $\frac{\sigma_{Z,ii}}{T}$ we get

$$\frac{\boldsymbol{z}_i' \boldsymbol{M}_D \left(\Delta \boldsymbol{y} - \boldsymbol{Z} \hat{\boldsymbol{\delta}} - \boldsymbol{W} \hat{\boldsymbol{\pi}} \right)}{T} - \frac{\lambda_G \sigma_{Z,ii} \hat{\delta}_i^s \left\| \hat{\boldsymbol{\delta}}^s \right\|_2^{-1}}{2T} - \frac{\lambda_{2,i} \sigma_{Z,ii} \operatorname{sign}(\hat{\delta}_i^s)}{2T} = 0 \quad (3.A.7)$$

The first term can be rewritten as

$$\frac{\boldsymbol{z}_i' \boldsymbol{M}_D \left(\Delta \boldsymbol{y} - \boldsymbol{Z} \hat{\boldsymbol{\delta}} - \boldsymbol{W} \hat{\boldsymbol{\pi}} \right)}{T} = \frac{\boldsymbol{z}_i' \boldsymbol{M}_D \left(\boldsymbol{\epsilon}_y - \boldsymbol{V} \boldsymbol{D}_T^{-1} \boldsymbol{D}_T (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \right)}{T} = O_p(1),$$

where the stochastic boundedness follows from the convergence in Lemma 3.A.2 and the result that $D_T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) = O_p(1)$ under the assumptions in Theorem 3.1. Regarding the second term in (3.A.7), note that $\hat{\delta}_i^s \left\| \hat{\boldsymbol{\delta}}^s \right\|_2^{-1} = O_p(1)$, because all estimates share the same rate of convergence. Then,

$$\frac{\lambda_G \sigma_{Z,ii} \hat{\delta}_i^s \left\| \hat{\boldsymbol{\delta}}^s \right\|_2^{-1}}{2T} \xrightarrow{p} 0,$$

since by our assumptions in Theorem 3.1, $\frac{\lambda_G \sigma_{Z,\max}}{\sqrt{T}} \to 0$. Finally, for the last term in

(3.A.7) we obtain

$$\frac{\lambda_{2,i}\sigma_{z,ii}}{2T} = \frac{\lambda_{\delta,T}\sigma_{Z,ii}}{2T\left|\hat{\delta}_{OLS,i}\right|^{k_\delta}} = \frac{\lambda_{\delta,T}\sigma_{Z,ii}}{T^{1-k_\delta}}\frac{1}{2\left|T\hat{\delta}_{OLS,i}\right|^{k_\delta}} \to \infty$$

under the assumption that $\frac{\lambda_{\delta,T}\sigma_{Z,\min}}{T^{1-k_\delta}} \to \infty$. This implies that

$$\mathbb{P}(\hat{\delta}_i^s = 0) = 1 - \mathbb{P}(\hat{\delta}_i^s \neq 0) \geq 1 - \mathbb{P}\left(\frac{\partial G_T(\gamma^s)}{\partial \delta_i^s}\Big|_{\gamma^s=\hat{\gamma}^s} = 0\right) \to 1. \tag{3.A.8}$$

Then, by noting that $\mathbb{P}(\hat{\delta}_i^s = 0) = \mathbb{P}(\hat{\delta}_i = 0)$, the selection consistency for $\hat{\delta}_i$ absent of cointegration follows.

Next, assume $\hat{\pi}_j^s \neq 0$ while $\pi_j = 0$ and let $\boldsymbol{w}_j$ be the j-th column of W. For $\hat{\pi}_j^s$ to be a minimum of (3.A.4) the first order conditions, after appropriate scaling, state

$$\frac{\boldsymbol{w}_j'\boldsymbol{M}_D\left(\Delta\boldsymbol{y} - \boldsymbol{V}\hat{\boldsymbol{\gamma}}\right)}{\sqrt{T}} - \frac{\lambda_{3,j}\sigma_{W,jj}\text{sign}(\hat{\pi}_j^s)}{2\sqrt{T}} = 0. \tag{3.A.9}$$

The first term can be rewritten as

$$\frac{\boldsymbol{w}_j'\boldsymbol{M}_D\left(\Delta\boldsymbol{y} - \boldsymbol{V}\hat{\boldsymbol{\gamma}}\right)}{\sqrt{T}} = \frac{\boldsymbol{w}_j'\boldsymbol{M}_D\left(\boldsymbol{\epsilon}_y - \boldsymbol{V}\boldsymbol{D}_T^{-1}\boldsymbol{D}_T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right)}{\sqrt{T}} = O_p(1),$$

where the stochastic boundedness follows from the Lemma 3.A.2 and $\boldsymbol{D}_T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) = O_p(1)$ by Theorem 3.1. For the second term in (3.A.9) we have

$$\frac{\lambda_{3,j}\sigma_{W,jj}}{2\sqrt{T}} = \frac{\lambda_{\pi,T}\sigma_{W,jj}}{2\sqrt{T}\left|\hat{\pi}_{OLS,j}\right|^{k_\pi}} = \frac{\lambda_{\pi,T}\sigma_{W,jj}}{T^{1/2-k_\pi/2}}\frac{1}{2\left|\sqrt{T}\hat{\pi}_{OLS,j}\right|^{k_\pi}} \to \infty \tag{3.A.10}$$

under the assumption that $\frac{\lambda_{\pi,T}\sigma_{W,\min}}{T^{1/2-k_\pi/2}} \to \infty$. The selection consistency for $\hat{\pi}_j$ then follows by the same argument used in (3.A.8).

The strategy for showing selection consistency in the presence of cointegration is analogous, albeit algebraically slightly more tedious. Let $\delta_i^s = \gamma_i^s = 0$. Then the first order condition for $\hat{\delta}_i^s \neq 0$ to be a minimum of the objective function, after pre-multiplying by $\frac{\sigma_{Z,ii}}{T}$, are again given by (3.A.7). Letting e_i denote the i-th column of I_{N+M}, the first term can be rewritten as

$$\begin{aligned}\frac{\boldsymbol{z}_i'\boldsymbol{M}_D\left(\boldsymbol{\epsilon}_y - \boldsymbol{V}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right)}{T} &= \frac{\boldsymbol{e}_i'\boldsymbol{V}'\boldsymbol{M}_D\left(\boldsymbol{\epsilon}_y - \boldsymbol{V}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right)}{T} \\ &= \frac{\boldsymbol{e}_i'\boldsymbol{Q}^{-1}\boldsymbol{S}_T}{T}\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}'\boldsymbol{M}_D\left(\boldsymbol{\epsilon}_y - \boldsymbol{V}\boldsymbol{Q}'\boldsymbol{S}_T^{-1}\boldsymbol{S}_T\boldsymbol{Q}'^{-1}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right) = O_p(1),\end{aligned}$$

because $\frac{\boldsymbol{e}_i'\boldsymbol{Q}^{-1}\boldsymbol{S}_T}{T} = O(1)$, $\boldsymbol{S}_T^{-1}QV'\boldsymbol{M}_D\boldsymbol{\epsilon}_y = O_p(1)$ and $\boldsymbol{S}_T^{-1}QV'\boldsymbol{M}_D\boldsymbol{V}\boldsymbol{Q}'\boldsymbol{S}_T^{-1} = O_p(1)$ by Lemma 3.A.3, and $\boldsymbol{S}_T\boldsymbol{Q}'^{-1}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) = O_p(1)$ by Theorem 3.1. The second term in (3.A.7) again converges to zero in probability and for the third and final term we obtain

$$\frac{\lambda_{2,i}\sigma_{Z,ii}}{2T} = \frac{\lambda_{\delta,T}\sigma_{Z,ii}}{2T\left|\hat{\delta}_{OLS,i}\right|^{k_\delta}} = \frac{\lambda_{\delta,T}\sigma_{Z,ii}}{T^{1-k_\delta/2}}\frac{1}{2\left|\sqrt{T}\hat{\delta}_{OLS,i}\right|^{k_\delta}} \to \infty,$$

under the assumption that $\frac{\lambda_{\delta,T}\sigma_{Z,ii}}{T^{1-k_\delta/2}} \to \infty$. Then, by the same argument as in (3.A.8) we can conclude that $\mathbb{P}(\hat{\delta}_i = 0) \to 1$.

Similarly, letting $\pi_j = 0$, the first order conditions for $\hat{\pi}_j \neq 0$ to be a minimum of (3.A.4) when $\pi_j = 0$ are again given by (3.A.9). The first term can be rewritten as

$$\frac{\boldsymbol{w}_j'\boldsymbol{M}_D\left(\boldsymbol{\epsilon}_y - \boldsymbol{V}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right)}{\sqrt{T}} = \frac{\boldsymbol{w}_j'\boldsymbol{M}_D\left(\boldsymbol{\epsilon}_y - \boldsymbol{V}\boldsymbol{Q}'\boldsymbol{S}_T^{-1}\boldsymbol{S}_T\boldsymbol{Q}'^{-1}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right)}{\sqrt{T}} = O_p(1),$$

because $\frac{\boldsymbol{w}_j'\boldsymbol{M}_D\boldsymbol{\epsilon}_y}{\sqrt{T}} = O_p(1)$ by Lemma 3.A.2,

$$\frac{\boldsymbol{w}_j'\boldsymbol{M}_D\boldsymbol{V}\boldsymbol{Q}'\boldsymbol{S}_T^{-1}}{\sqrt{T}} = \begin{bmatrix} T^{-1}\boldsymbol{w}_j'\boldsymbol{Z}_{-1}\boldsymbol{B} & T^{-1}\boldsymbol{w}_j'\boldsymbol{W} & T^{-3/2}\boldsymbol{w}_j'\boldsymbol{Z}_{-1}\boldsymbol{A}_\perp \end{bmatrix} = O_p(1),$$

by Lemma 3.A.3 and $\boldsymbol{S}_T\boldsymbol{Q}'^{-1}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) = O_p(1)$ by Theorem 3.1. Furthermore, we again have that $\frac{\lambda_{3,j}\sigma_{W,jj}}{2\sqrt{T}} \to \infty$ based on (3.A.10). Consequently, it follows that $\mathbb{P}(\hat{\pi}_j = 0) \to 1$ by the same argument used for (3.A.8), thus completing the proof. ■

Proof of Theorem 3.3. Without loss of generality we impose an ordering on the variables such that $\boldsymbol{V} = (\boldsymbol{V}_{S_\gamma}, \boldsymbol{V}_{S_\gamma^c}) = (\boldsymbol{Z}_{S_\delta}, \boldsymbol{W}_{S_\pi}, \boldsymbol{Z}_{S_\delta^c}, \boldsymbol{W}_{S_\pi^c})$, where the variables collected in $\boldsymbol{V}_{S_\gamma}$ carry non-zero coefficients in the true DGP, whereas $\boldsymbol{V}_{S_\gamma^c}$ contains all irrelevant variables. The de-standardized estimate $\hat{\boldsymbol{\gamma}}_{S_\gamma}$ is defined as $\sigma_{V_{S_\gamma}}^{-1}\hat{\boldsymbol{\gamma}}_{S_\gamma}^s$. Since $\hat{\boldsymbol{\gamma}}^s$ are the minimizers of (3.A.4), they must set the subgradient equations equal to zero:

$$\tilde{\boldsymbol{V}}'\boldsymbol{M}_D(\Delta\boldsymbol{y} - \tilde{\boldsymbol{V}}\hat{\boldsymbol{\gamma}}^s) - \frac{1}{2}\hat{s}\left(\hat{\gamma}^s\right) = 0,$$

or after pre-multiplication with $\boldsymbol{\Sigma}_V$ by

$$\boldsymbol{V}'\boldsymbol{M}_D(\Delta\boldsymbol{y} - \boldsymbol{V}\hat{\boldsymbol{\gamma}}) - \frac{1}{2}\boldsymbol{\Sigma}_V\hat{s}\left(\hat{\gamma}^s\right) = 0, \tag{3.A.11}$$

where we let $\hat{s}\left(\hat{\boldsymbol{\gamma}}^s\right)$ denote the sub-gradient of the penalty function $P_\lambda(\hat{\boldsymbol{\gamma}}^s)$. In particular, define $\Lambda = \text{diag}(\lambda_2, \lambda_3)$, then

$$\hat{s}\left(\hat{\boldsymbol{\gamma}}^s\right) = \lambda_G \hat{s}_G(\hat{\boldsymbol{\delta}}^s) + \Lambda \hat{s}_I\left(\hat{\boldsymbol{\gamma}}^s\right),$$

where $\hat{s}_G(\hat{\boldsymbol{\delta}}^s)$ is a $(N+M)$-dimensional vector with the first N elements being given by $\hat{\boldsymbol{\delta}} / \left\|\hat{\boldsymbol{\delta}}\right\|_2$, whenever at least one of the $\hat{\delta}_j \neq 0$, or by a N-dimensional vector x with $\|x\|_2 \leq 1$ otherwise, and the remaining M elements of $\hat{s}_G(\hat{\boldsymbol{\delta}}^s)$ are equal to zero. Furthermore, $\hat{s}_I\left(\hat{\boldsymbol{\gamma}}^s\right)$, has element j equal to $\text{sign}(\hat{\gamma}_j^s)$ when $\hat{\gamma}_j^s \neq 0$ and can be any scalar $x \in [-1, 1]$ otherwise. Below we will additionally refer to the vector

$$\hat{s}\left(\hat{\boldsymbol{\gamma}}^s_{S_\gamma}\right) = \lambda_G \hat{s}_G(\hat{\boldsymbol{\gamma}}^s_{S_\gamma}) + \Lambda_{S_\gamma} \hat{s}_I\left(\hat{\boldsymbol{\gamma}}^s_{S_\gamma}\right),$$

which is the sub-gradient of the penalty function for the coefficients indexed by S_γ. Important to note is that given our assumptions on the penalty terms, i.e. $\frac{\lambda_{G,T}\sigma_{Z,\max}}{\sqrt{T}} \xrightarrow{p} 0$, $\frac{\lambda_{\delta,T}\sigma_{Z,\max}}{\sqrt{T}} \to 0$ and $\frac{\lambda_{\pi,T}\sigma_{W,\max}}{\sqrt{T}} \to 0$, it immediately follows that

$$T^{-1/2}\boldsymbol{\Sigma}_{V,S_\gamma}\hat{s}\left(\hat{\boldsymbol{\gamma}}^s_{S_\gamma}\right) \to 0. \tag{3.A.12}$$

We proceed by rewriting the first order conditions (3.A.11) in terms of $\hat{\boldsymbol{\gamma}}_{S_\gamma}$ as

$$\begin{aligned}
\mathbf{0} =& \boldsymbol{V}'_{S_\gamma}\boldsymbol{M}_D\left(\Delta\boldsymbol{y} - \boldsymbol{V}_{S_\gamma}\hat{\gamma}_{S_\gamma} - \boldsymbol{V}_{S^c_\gamma}\hat{\boldsymbol{\gamma}}_{S^c_\gamma}\right) - \frac{1}{2}\boldsymbol{\Sigma}_{V,S_\gamma}\hat{s}\left(\hat{\boldsymbol{\gamma}}^s_{S_\gamma}\right) \\
=& \boldsymbol{V}'_{S_\gamma}\boldsymbol{M}_D\left(\hat{\boldsymbol{\epsilon}}_{OLS} - \boldsymbol{V}_{S_\gamma}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \hat{\boldsymbol{\gamma}}_{OLS,S_\gamma}\right) - \boldsymbol{V}_{S^c_\gamma}\hat{\boldsymbol{\gamma}}_{S^c_\gamma}\right) - \frac{1}{2}\boldsymbol{\Sigma}_{V,S_\gamma}\hat{s}\left(\hat{\boldsymbol{\gamma}}^s_{S_\gamma}\right) \\
=& -\boldsymbol{V}'_{S_\gamma}\boldsymbol{M}_D\left(\boldsymbol{V}_{S_\gamma}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \hat{\boldsymbol{\gamma}}_{OLS,S_\gamma}\right) + \boldsymbol{V}_{S^c_\gamma}\hat{\boldsymbol{\gamma}}_{S^c_\gamma}\right) - \frac{1}{2}\boldsymbol{\Sigma}_{V,S_\gamma}\hat{s}\left(\hat{\boldsymbol{\gamma}}^s_{S_\gamma}\right),
\end{aligned} \tag{3.A.13}$$

where $\hat{\boldsymbol{\epsilon}}_{OLS} = \boldsymbol{M}_D(\Delta\boldsymbol{y} - \boldsymbol{V}_{S_\gamma}\hat{\boldsymbol{\gamma}}_{OLS,S_\gamma})$ such that $\boldsymbol{V}'_{S_\gamma}\boldsymbol{M}_D\hat{\boldsymbol{\epsilon}}_{OLS} = \mathbf{0}$ by construction. Reordering terms in (3.A.13) gives

$$\begin{aligned}
\hat{\boldsymbol{\gamma}}_{S_\gamma} - \hat{\boldsymbol{\gamma}}_{OLS,S_\gamma} =& \left(\boldsymbol{V}'_{S_\gamma}\boldsymbol{M}_D\boldsymbol{V}_{S_\gamma}\right)^{-1}\boldsymbol{V}'_{S_\gamma}\boldsymbol{M}_D\boldsymbol{V}_{S^c_\gamma}\hat{\boldsymbol{\gamma}}_{S^c_\gamma} \\
&- \frac{1}{2}\left(\boldsymbol{V}'_{S_\gamma}\boldsymbol{M}_D\boldsymbol{V}_{S_\gamma}\right)^{-1}\boldsymbol{\Sigma}_{V,S_\gamma}\hat{s}\left(\hat{\boldsymbol{\gamma}}^s_{S_\gamma}\right).
\end{aligned} \tag{3.A.14}$$

We now separately consider the cases without and with cointegration in the underlying DGP. Absent of cointegration we have $\boldsymbol{V}_{S_\gamma} = \boldsymbol{W}_{S_\pi}$ and $\boldsymbol{\gamma}_{S_\gamma} = \boldsymbol{\pi}_{S_\pi}$ such that after

appropriately scaling (3.A.14) we obtain

$$\sqrt{T}\left(\hat{\boldsymbol{\pi}}_{S_\pi} - \hat{\boldsymbol{\pi}}_{OLS,S_\pi}\right) = -\left(T^{-1}\boldsymbol{W}'_{S_\pi}\boldsymbol{M}_D\boldsymbol{W}_{S_\pi}\right)^{-1} T^{-1/2}\boldsymbol{W}'_{S_\pi}\boldsymbol{M}_D\boldsymbol{V}_{S^c_\gamma}\hat{\boldsymbol{\gamma}}_{S^c_\gamma} - \frac{1}{2}\left(T^{-1}\boldsymbol{W}'_{S_\pi}\boldsymbol{M}_D\boldsymbol{W}_{S_\pi}\right)^{-1} T^{-1/2}\boldsymbol{\Sigma}_{W,S_\pi}\hat{s}\left(\hat{\boldsymbol{\gamma}}^s_{S_\gamma}\right) = o_p(1),$$

where the stated convergence follows, because $\mathbb{P}(\hat{\gamma}_{S^c_\gamma,i} = 0) \to 1$, for all $i \in S^c_\gamma$, such that

$$\left(T^{-1}\boldsymbol{W}'_{S_\pi}\boldsymbol{M}_D\boldsymbol{W}_{S_\pi}\right)^{-1} T^{-1/2}\boldsymbol{W}'_{S_\pi}\boldsymbol{M}_D\boldsymbol{V}_{S^c_\gamma}\hat{\boldsymbol{\gamma}}_{S^c_\gamma}$$

vanishes in probability and

$$\frac{1}{2}\left(T^{-1}\boldsymbol{W}'_{S_\pi}\boldsymbol{M}_D\boldsymbol{W}_{S_\pi}\right)^{-1} T^{-1/2}\boldsymbol{\Sigma}_{W,S_\pi}\hat{s}\left(\hat{\boldsymbol{\gamma}}^s_{S_\gamma}\right) = o_p(1)$$

by Lemma 3.A.2 and (3.A.12). Alternatively, when cointegration is present in the data we make use of $\boldsymbol{S}_{T,S_\gamma}$ and $\boldsymbol{Q}_{S_\gamma}$ as defined in Theorem 3.3. Observe that

$$\boldsymbol{Q}_{S_\gamma}\boldsymbol{v}_{S_\gamma,t} = \begin{bmatrix} \boldsymbol{B}'_{S_\delta}\boldsymbol{z}_{S_\delta,t-1} \\ \boldsymbol{w}_{S_\pi,t} \\ \boldsymbol{B}'_{S_\delta,\perp}\boldsymbol{z}_{S_\delta,t-1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_{S_{\gamma_1},t} \\ \boldsymbol{v}_{S_{\gamma_2},t} \end{bmatrix},$$

where $\boldsymbol{v}_{S_{\gamma_1},t} = (\boldsymbol{z}'_{S_\delta,t-1}\boldsymbol{B}_{S_\delta}, \boldsymbol{w}'_t)' \sim I(0)$ and $\boldsymbol{v}_{S_{\gamma_2},t} = \boldsymbol{z}_{S_\delta,t-1}\boldsymbol{B}_{\perp,S_\delta} \sim I(1)$. In matrix form, we write

$$\boldsymbol{V}_{s_\gamma}\boldsymbol{Q}'_{S_\gamma} = \begin{bmatrix} \boldsymbol{V}_{S_\gamma,1} & \boldsymbol{V}_{S_\gamma,2} \end{bmatrix},$$

with $\boldsymbol{V}_{S_\gamma,1} = \begin{bmatrix} \boldsymbol{Z}_{-1,S_\delta}\boldsymbol{B}_{S_\delta} & \boldsymbol{W}_{S_\pi} \end{bmatrix}$ and $\boldsymbol{V}_{S_\gamma,2} = \boldsymbol{Z}_{-1,S_\delta}\boldsymbol{B}_{S_\delta,\perp}$. By a straightforward adaptation[17] of Lemma 3, it then follows that

$$\begin{aligned} &\boldsymbol{S}^{-1}_{T,S_\gamma}\boldsymbol{Q}_{S_\gamma}\boldsymbol{V}'_{S_\gamma}\boldsymbol{M}_D\boldsymbol{V}_{S_\gamma}\boldsymbol{Q}'_{S_\gamma}\boldsymbol{S}^{-1}_{T,S_\gamma} \\ &\quad\xrightarrow{d} \begin{bmatrix} \boldsymbol{\Sigma}_{V_{S_\gamma,1}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}'_{S_\delta,\perp}\boldsymbol{C}_{S_\delta}\left(\int_0^1 \boldsymbol{B}^*_{S_\delta}(r)\boldsymbol{B}^*_{S_\delta}(r)'\right)\boldsymbol{C}'_{S_\delta}\boldsymbol{B}_{S_\delta,\perp} \end{bmatrix}, \end{aligned} \tag{3.A.15}$$

where

$$\boldsymbol{\Sigma}_{V_{S_\gamma,1}} = \begin{bmatrix} \mathbb{E}\left(\boldsymbol{B}'_{S_\delta}\boldsymbol{u}_{S_\delta,t}\boldsymbol{u}'_{S_\delta,t}\boldsymbol{B}_{S_\delta}\right) & \boldsymbol{0} \\ \boldsymbol{0} & \mathbb{E}\left(\boldsymbol{w}_{S_\pi,t}\boldsymbol{w}'_{S_\pi,t}\right) \end{bmatrix},$$

[17] The adaptation of Lemma 3.A.3 follows from replacing $\boldsymbol{A}_\perp$, $\boldsymbol{\Sigma}_{V_1}$ and $\boldsymbol{C}$ with $\boldsymbol{B}_{S_\delta,\perp}$, $\boldsymbol{\Sigma}_{V_{S_\gamma,1}}$ and $\boldsymbol{C}_{S_\delta}$, respectively.

and $\boldsymbol{C}_{S_\delta} = \boldsymbol{B}_{\perp,S_\delta} \left(\boldsymbol{A}'_\perp \boldsymbol{B}_\perp\right)^{-1}$. Then, it follows that

$$\begin{aligned}
&\boldsymbol{S}_{T,S_\gamma} \boldsymbol{Q}'^{-1}_{S_\gamma} \left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \hat{\boldsymbol{\gamma}}_{OLS,S_\gamma}\right) \\
&\quad = -\left(\boldsymbol{S}^{-1}_{T,S_\gamma} \boldsymbol{Q}_{S_\gamma} \boldsymbol{V}'_{S_\gamma} \boldsymbol{M}_D \boldsymbol{V}_{S_\gamma} \boldsymbol{Q}'_{S_\gamma} \boldsymbol{S}^{-1}_{T,S_\gamma}\right)^{-1} \boldsymbol{S}^{-1}_{T,S_\gamma} \boldsymbol{Q}_{S_\gamma} \boldsymbol{V}'_{S_\gamma} \boldsymbol{M}_D \boldsymbol{V}_{S^c_\gamma} \hat{\boldsymbol{\gamma}}_{S^c_\gamma} \\
&- \frac{1}{2}\left(\boldsymbol{S}^{-1}_{T,S_\gamma} \boldsymbol{Q}_{S_\gamma} \boldsymbol{V}'_{S_\gamma} \boldsymbol{M}_D \boldsymbol{V}_{S_\gamma} \boldsymbol{Q}'_{S_\gamma} \boldsymbol{S}^{-1}_{T,S_\gamma}\right)^{-1} \boldsymbol{S}^{-1}_{T,S_\gamma} \boldsymbol{Q}_{S_\gamma} \boldsymbol{\Sigma}_{V,S_\gamma} \hat{s}\left(\hat{\boldsymbol{\gamma}}^s_{S_\gamma}\right) = o_p(1),
\end{aligned}$$

where the convergence follows because

$$\left(\boldsymbol{S}^{-1}_{T,S_\gamma} \boldsymbol{Q}_{S_\gamma} \boldsymbol{V}'_{S_\gamma} \boldsymbol{M}_D \boldsymbol{V}_{S_\gamma} \boldsymbol{Q}'_{S_\gamma} \boldsymbol{S}^{-1}_{T,S_\gamma}\right)^{-1} \boldsymbol{S}^{-1}_{T,S_\gamma} \boldsymbol{Q}_{S_\gamma} \boldsymbol{V}'_{S_\gamma} \boldsymbol{M}_D \boldsymbol{V}_{S^c_\gamma} \hat{\boldsymbol{\gamma}}_{S^c_\gamma}$$

vanishes in probability since $\mathbb{P}(\hat{\gamma}_{S^c_\gamma,i} = 0) \to 1$, for all $i \in S^c_\gamma$, and

$$\frac{1}{2}\left(\boldsymbol{S}^{-1}_{T,S_\gamma} \boldsymbol{Q}_{S_\gamma} \boldsymbol{V}'_{S_\gamma} \boldsymbol{M}_D \boldsymbol{V}_{S_\gamma} \boldsymbol{Q}'_{S_\gamma} \boldsymbol{S}^{-1}_{T,S_\gamma}\right)^{-1} \boldsymbol{S}^{-1}_{T,S_\gamma} \boldsymbol{Q}_{S_\gamma} \boldsymbol{\Sigma}_{V,S_\gamma} \hat{s}\left(\hat{\boldsymbol{\gamma}}^s_{S_\gamma}\right) = o_p(1)$$

because of (3.A.15) and (3.A.12). This completes the proof. ■

Proof of Corollary 3.1. We first show that

$$\begin{aligned}
&\boldsymbol{S}_{T,S_\gamma} \boldsymbol{Q}'^{-1}_{S_\gamma} (\hat{\boldsymbol{\gamma}}_{S_\gamma,OLS} - \boldsymbol{\gamma}_{S_\gamma}) \xrightarrow{d} \\
&\begin{bmatrix} \mathcal{N}\left(\boldsymbol{0}, \sigma^2_{\epsilon_y} \boldsymbol{\Sigma}^{-1}_{V_{S_\gamma,1}}\right) \\ \left(\boldsymbol{B}'_{S_\delta,\perp} \boldsymbol{C}_{S_\delta} \left(\int_0^1 \boldsymbol{B}^*_{S_\delta}(r) \boldsymbol{B}^*_{S_\delta}(r)'\right) \boldsymbol{C}'_{S_\delta} \boldsymbol{B}_{S_\delta,\perp}\right)^{-1} \boldsymbol{B}'_{S_\delta,\perp} \boldsymbol{C}_{S_\delta} \left(\int_0^1 \boldsymbol{B}^*_{S_\delta}(r) dB_\epsilon(r)\right) \end{bmatrix},
\end{aligned} \tag{3.A.16}$$

where $\sigma^2_{\epsilon_y} = \mathbb{E}(\epsilon^2_{y,t})$. Note that

$$\begin{aligned}
&\boldsymbol{S}_{T,S_\gamma} \boldsymbol{Q}'^{-1}_{S_\gamma} (\hat{\boldsymbol{\gamma}}_{S_\gamma,OLS} - \boldsymbol{\gamma}_{S_\gamma}) = \boldsymbol{S}_{TS_\gamma} \boldsymbol{Q}'^{-1}_{S_\gamma} \left(\boldsymbol{V}'_{S_\gamma} \boldsymbol{M}_D \boldsymbol{V}_{S_\gamma}\right)^{-1} \boldsymbol{V}'_{S_\gamma} \boldsymbol{M}_D \boldsymbol{\epsilon}_y \\
&= \left(\boldsymbol{S}^{-1}_{T,S_\gamma} \boldsymbol{Q}_{S_\gamma} \boldsymbol{V}'_{S_\gamma} \boldsymbol{M}_D \boldsymbol{V}_{S_\gamma} \boldsymbol{Q}' \boldsymbol{S}^{-1}_{T,S_\gamma}\right)^{-1} \boldsymbol{S}^{-1}_{T,S_\gamma} \boldsymbol{Q}_{S_\gamma} \boldsymbol{V}'_{S_\gamma} \boldsymbol{M}_D \boldsymbol{\epsilon}_y.
\end{aligned}$$

By a straightforward adaptation of Lemma 3.A.3 it follows that

$$\boldsymbol{S}^{-1}_{T,S_\gamma} \boldsymbol{Q}_{S_\gamma} \boldsymbol{V}'_{S_\gamma} \boldsymbol{M}_D \boldsymbol{\epsilon}_y \xrightarrow{d} \begin{bmatrix} N\left(\boldsymbol{0}, \sigma^2_{\epsilon_y} \boldsymbol{\Sigma}_{V_{S_\gamma,1}}\right) \\ \boldsymbol{B}'_{S_\delta,\perp} \boldsymbol{C}_{S_\delta} \left(\int_0^1 \boldsymbol{B}^*_{S_\delta}(r) dB_\epsilon(r)\right) \end{bmatrix},$$

such that by (3.A.15) in combination with the continuous mapping theorem for functionals and Slutsky's theorem, the result in (3.A.16) follows.

As a direct consequence, we have

$$T^{1/2}\boldsymbol{Q}'^{-1}_{S_\gamma}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma}-\boldsymbol{\gamma}_{S_\gamma}\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\sigma^2_{\epsilon_y}\begin{bmatrix}\boldsymbol{\Sigma}^{-1}_{V_{S_\gamma,1}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\right),$$

such that

$$\begin{aligned}\sqrt{T}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma}-\boldsymbol{\gamma}_{S_\gamma}\right)&\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\sigma^2_{\epsilon_y}\boldsymbol{Q}'_{S_\gamma}\begin{bmatrix}\boldsymbol{\Sigma}^{-1}_{V_{S_\gamma,1}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\boldsymbol{Q}_{S_\gamma}\right)\\&=\mathcal{N}\left(\mathbf{0},\sigma^2_{\epsilon_y}\begin{bmatrix}\boldsymbol{B}_{S_\delta}\boldsymbol{\Sigma}_U^{-1}\boldsymbol{B}'_{S_\delta} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}^{-1}_{W_{S_\pi}}\end{bmatrix}\right),\end{aligned}$$

with $\boldsymbol{\Sigma}_U$ and $\boldsymbol{\Sigma}_{W_{S_\pi}}$ as defined in Corollary 3.1. This proves the part of Corollary 3.1 on the convergence of the estimator.

We proceed by showing that the matrix $\boldsymbol{B}_{S_\delta}\boldsymbol{\Sigma}_U^{-1}\beta'_{S_\delta}$ is uniquely defined, regardless of the choice of the basis matrix $\boldsymbol{B}_{S_\delta}$. Naturally, the basis matrix $\boldsymbol{B}_{S_\delta}$ itself is not unique, as any matrix whose columns form a basis for the left nullspace of $\boldsymbol{B}_{\perp,S_\delta}$ may be used in the construction of $\boldsymbol{Q}_{S_\gamma}$. Accordingly, assume that another matrix satisfying this condition is given by $\boldsymbol{B}^*_{S_\delta}$ with the i-th column vector given by $\beta^*_{S_\delta,i}=\boldsymbol{B}_{S_\delta}\boldsymbol{x}_i$, where $\boldsymbol{x}_i$ are the coordinates of $\beta^*_{S_\delta,i}$ with respect to the basis $\boldsymbol{B}_{S_\delta}$. Then, we can represent our new basis as

$$\boldsymbol{B}^*_{S_\delta}=\boldsymbol{B}_{S_\delta}\boldsymbol{X},$$

where $\boldsymbol{X}=\begin{bmatrix}\boldsymbol{x}_1 & \dots & \boldsymbol{x}_{r_2}\end{bmatrix}$. Moreover, $\boldsymbol{X}$ must be linearly independent, because otherwise there exists a $\boldsymbol{u}\in R^{r_2}$ with $\boldsymbol{u}\neq\mathbf{0}$ and

$$\boldsymbol{B}^*_{S_\delta}\boldsymbol{u}=\boldsymbol{B}_{S_\delta}\boldsymbol{X}\boldsymbol{u}=\mathbf{0},$$

thereby contradicting the claim that $\boldsymbol{B}^*_{S_\delta}$ is a basis matrix. Consequently, $\boldsymbol{X}$ is an invertible linear transformation and it follows that

$$\begin{aligned}\boldsymbol{B}^*_{S_\delta}\boldsymbol{\Sigma}^{*-1}_U\boldsymbol{B}^{*\prime}_{S_\delta}&=\boldsymbol{B}^*_{S_\delta}\left(\mathbb{E}\left(\boldsymbol{B}^{*\prime}_{S_\delta}\boldsymbol{u}_{S_\delta,t}\boldsymbol{u}'_{S_\delta,t}\boldsymbol{B}^*_{S_\delta}\right)\right)^{-1}\boldsymbol{B}^{*\prime}_{S_\delta}\\&=\boldsymbol{B}_{S_\delta}\boldsymbol{X}\left(\mathbb{E}\left(\boldsymbol{X}'\boldsymbol{B}'_{S_\delta}\boldsymbol{u}_{S_\delta,t}\boldsymbol{u}'_{S_\delta,t}\boldsymbol{B}_{S_\delta}\boldsymbol{X}\right)\right)^{-1}\boldsymbol{X}'\boldsymbol{B}'_{S_\delta}\\&=\boldsymbol{B}_{S_\delta}\left(\mathbb{E}\left(\boldsymbol{B}'_{S_\delta}\boldsymbol{u}_{S_\delta,t}\boldsymbol{u}'_{S_\delta,t}\boldsymbol{B}_{S_\delta}\right)\right)^{-1}\boldsymbol{B}'_{S_\delta}=\boldsymbol{B}_{S_\delta}\boldsymbol{\Sigma}_U^{-1}\boldsymbol{B}'_{S_\delta},\end{aligned}$$

thereby validating the claim that $\boldsymbol{B}_{S_\delta}\boldsymbol{\Sigma}_U^{-1}\boldsymbol{B}'_{S_\delta}$ is uniquely defined regardless of the choice of basis. ■

Appendix 3.B Supplementary Material

3.B.1 Proof of Corollary 3.2

Proof of Corollary 3.2. The proof of the consistency of the estimated deterministic components is straightforward, though algebraically tedious. Recall that $\boldsymbol{\theta} = (\mu_0, \tau_0)'$. Based on Lemma 3.A.4 it follows that

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= (\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'(\Delta\boldsymbol{y} - \boldsymbol{V}\hat{\boldsymbol{\gamma}}) = (\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'\left(\hat{\boldsymbol{\epsilon}}_{y,OLS} - \boldsymbol{V}(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}_{OLS}) + \boldsymbol{D}\hat{\boldsymbol{\theta}}_{OLS}\right) \\ &= \hat{\boldsymbol{\theta}}_{OLS} - (\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{V}(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}_{OLS}),\end{aligned}$$

such that

$$\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS} = -(\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{V}(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}_{OLS}). \tag{3.B.1}$$

Note that

$$(\boldsymbol{D}'\boldsymbol{D})^{-1} = \frac{1}{|\boldsymbol{D}'\boldsymbol{D}|}\begin{bmatrix}\bar{\boldsymbol{t}}'\bar{\boldsymbol{t}} & -\boldsymbol{\iota}'\bar{\boldsymbol{t}} \\ -\boldsymbol{\iota}'\bar{\boldsymbol{t}} & T\end{bmatrix},$$

where

$$|\boldsymbol{D}'\boldsymbol{D}| = T\bar{\boldsymbol{t}}'\bar{\boldsymbol{t}} - (\boldsymbol{\iota}'\bar{\boldsymbol{t}})^2 = O(T^4).$$

The analytical expression for the constant can be derived from (3.B.1). Assuming for the moment that $\boldsymbol{\mu} \neq \boldsymbol{0}$, $\boldsymbol{\tau} \neq \boldsymbol{0}$ and $\boldsymbol{\delta} = \boldsymbol{0}$, we obtain

$$\begin{aligned}\hat{\mu}_0 - \hat{\mu}_{0,OLS} &= \frac{1}{|\boldsymbol{D}'\boldsymbol{D}|}\left[(\bar{\boldsymbol{t}}'\bar{\boldsymbol{t}}\boldsymbol{\iota}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\bar{\boldsymbol{t}}')\boldsymbol{V}\right]\left[\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}_{OLS}\right] \\ &= \frac{1}{|\boldsymbol{D}'\boldsymbol{D}|}\left[(\bar{\boldsymbol{t}}'\bar{\boldsymbol{t}}\boldsymbol{\iota}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\bar{\boldsymbol{t}}')\boldsymbol{Z}_{-1} \quad (\bar{\boldsymbol{t}}'\bar{\boldsymbol{t}}\boldsymbol{\iota}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\bar{\boldsymbol{t}}')\boldsymbol{W}\right]\begin{bmatrix}\hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}}_{OLS} \\ \hat{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}}_{OLS}\end{bmatrix} \\ &= O(T^{-4})\begin{bmatrix}O_p(T^{9/2}) & O_p(T^4)\end{bmatrix}\begin{bmatrix}o_p(T^{-1}) \\ o_p(T^{-1/2})\end{bmatrix} = o_p(T^{-1/2}).\end{aligned} \tag{3.B.2}$$

This may be verified by writing out each term and applying Lemma 3.A.2. We demonstrate this for this particular instance. Note that

$$\begin{aligned}(\bar{\boldsymbol{t}}'\bar{\boldsymbol{t}}\boldsymbol{\iota}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\bar{\boldsymbol{t}}')\boldsymbol{Z}_{-1} &= (\bar{\boldsymbol{t}}'\bar{\boldsymbol{t}}\boldsymbol{\iota}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\bar{\boldsymbol{t}}')(\boldsymbol{S}_{-1}\boldsymbol{C}' + \boldsymbol{\iota}\boldsymbol{\mu}' + \bar{\boldsymbol{t}}\boldsymbol{\tau}' + \boldsymbol{U}_{-1}) \\ &= (\bar{\boldsymbol{t}}'\bar{\boldsymbol{t}}\boldsymbol{\iota}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\bar{\boldsymbol{t}}')\boldsymbol{S}_{-1}\boldsymbol{C}' + \left(T\bar{\boldsymbol{t}}'\bar{\boldsymbol{t}} - (\boldsymbol{\iota}'\bar{\boldsymbol{t}})^2\right)\boldsymbol{\mu}' + (\bar{\boldsymbol{t}}'\bar{\boldsymbol{t}}\boldsymbol{\iota}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\bar{\boldsymbol{t}}')\boldsymbol{U}_{-1} \\ &= O_p(T^{9/2}) + O(T^4) + O_p(T^{7/2}).\end{aligned}$$

Hence, regardless of whether $\boldsymbol{\mu} \neq \mathbf{0}$ or $\boldsymbol{\tau} \neq \mathbf{0}$, it holds that $(\bar{\boldsymbol{t}}'\bar{\boldsymbol{t}}\boldsymbol{\iota}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\bar{\boldsymbol{t}}')\,\boldsymbol{Z}_{-1} = O_p(T^{9/2})$. Similarly, for the term in (3.B.2) involving $\boldsymbol{W}$, we note that

$$\boldsymbol{W} = \begin{bmatrix}\Delta\boldsymbol{X} & \Delta\boldsymbol{Z}_{-1} & \dots & \Delta\boldsymbol{Z}_{-p}\end{bmatrix} = \begin{bmatrix}\Delta\boldsymbol{Z} & \dots & \Delta\boldsymbol{Z}_{-p}\end{bmatrix}\begin{bmatrix}\mathbf{0}_{1\times((P+1)N-1)} \\ I_{(P+1)N-1}\end{bmatrix},$$

where $\Delta\boldsymbol{Z}_{-j} = \boldsymbol{\iota}\boldsymbol{\tau}' + \boldsymbol{U}_{-j}$ with

$$\boldsymbol{U}'_{-j} = (\boldsymbol{C} + \boldsymbol{C}(L)(1-L))\begin{bmatrix}\mathbf{0}_{N\times j} & \boldsymbol{\epsilon}_1 & \dots & \boldsymbol{\epsilon}_{T-j}\end{bmatrix}.$$

Then, since

$$(\bar{\boldsymbol{t}}'\bar{\boldsymbol{t}}\boldsymbol{\iota}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\bar{\boldsymbol{t}}')\,\Delta\boldsymbol{Z}_{-j} = (\bar{\boldsymbol{t}}'\bar{\boldsymbol{t}}\boldsymbol{\iota}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\bar{\boldsymbol{t}}')\,\boldsymbol{\iota}\boldsymbol{\tau}' + (\bar{\boldsymbol{t}}'\bar{\boldsymbol{t}}\boldsymbol{\iota}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\bar{\boldsymbol{t}}')\,\boldsymbol{U}_j = O(T^4) + O_p(T^{7/2}),$$

it follows that $\boldsymbol{W} = O_p(T^4)$ when $\boldsymbol{\tau} \neq \mathbf{0}$ and $\boldsymbol{W} = O_p(T^{7/2})$ when $\boldsymbol{\tau} = \mathbf{0}$. However, when $\boldsymbol{\tau} = \mathbf{0}$ the rate of $\hat{\mu}_0$ will be determined by the term in (3.B.2) involving $\boldsymbol{Z}_{-1}$ and the convergence rate is thus invariant to the presence of a constant or deterministic trend.

In the remainder of the proof we proceed along a similar strategy by deriving the stochastic order for varying $\boldsymbol{\delta}$, $\boldsymbol{\tau}$ and $\boldsymbol{\mu}$. However, for the sake of brevity, we refrain from writing out each individual term and rather refer to each term's stochastic order directly. We start by deriving a similar result to (3.B.2), but for the case $\boldsymbol{\delta} \neq \mathbf{0}$. Then, (3.B.1) can be written as

$$\begin{aligned}\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS} &= -\left(\boldsymbol{D}'\boldsymbol{D}\right)^{-1}\boldsymbol{D}'\boldsymbol{V}\boldsymbol{Q}'\boldsymbol{Q}'^{-1}\left(\hat{\gamma} - \hat{\gamma}_{OLS}\right) \\ &= -\left(\boldsymbol{D}'\boldsymbol{D}\right)^{-1}\boldsymbol{D}'\begin{bmatrix}\boldsymbol{Z}_{-1}\boldsymbol{B} & \boldsymbol{W} & \boldsymbol{Z}_{-1}\boldsymbol{A}_\perp\end{bmatrix}\begin{bmatrix}(\boldsymbol{A}'\boldsymbol{B})^{-1}\boldsymbol{A}'(\hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}}_{OLS}) \\ \hat{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}}_{OLS} \\ (\boldsymbol{B}'_\perp\boldsymbol{A}_\perp)^{-1}\boldsymbol{B}'_\perp(\hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}}_{OLS})\end{bmatrix},\end{aligned} \tag{3.B.3}$$

from which follows that,

$$\begin{aligned}\hat{\mu}_0 - \hat{\mu}_{0,OLS} =& \frac{1}{|\boldsymbol{D}'\boldsymbol{D}|}\left[(\bar{\boldsymbol{t}}'\bar{\boldsymbol{t}}\boldsymbol{\iota}' - \boldsymbol{\iota}'\bar{b}\bar{t}\bar{b}\bar{t}')\boldsymbol{Z}_{-1}\boldsymbol{B} \quad (\bar{\boldsymbol{t}}'\bar{\boldsymbol{t}}\boldsymbol{\iota}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\bar{\boldsymbol{t}}')\boldsymbol{W} \quad (\bar{\boldsymbol{t}}'\bar{\boldsymbol{t}}\boldsymbol{\iota}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\bar{\boldsymbol{t}}')\boldsymbol{Z}_{-1}\boldsymbol{A}_\perp\right] \\ &\times \begin{bmatrix} (\boldsymbol{A}'\boldsymbol{B})^{-1}\boldsymbol{A}'(\hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}}_{OLS}) \\ \hat{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}}_{OLS} \\ (\boldsymbol{B}_\perp'\boldsymbol{A}_\perp)^{-1}\boldsymbol{B}_\perp'(\hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}}_{OLS}) \end{bmatrix} \\ =& O(T^{-4})\left[O_p(T^4) \quad O_p(T^4) \quad O_p(T^{9/2})\right]\begin{bmatrix} o_p(T^{-1/2}) \\ o_p(T^{-1/2}) \\ o_p(T^{-1}) \end{bmatrix} = o_p(T^{-1/2}).\end{aligned}$$

Again, one may verify that the rate of convergence holds irrespective of whether $\boldsymbol{\mu} = \boldsymbol{0}$ or $\boldsymbol{\tau} = \boldsymbol{0}$.

Next, we move on to the expression for the trend coefficient. For the cases with $\boldsymbol{B} = \boldsymbol{0}$, we will rely on the expression

$$\hat{\tau}_0 - \hat{\tau}_{0,OLS} = \frac{1}{|\boldsymbol{D}'\boldsymbol{D}|}\left[(T\bar{\boldsymbol{t}}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\boldsymbol{\iota}')\boldsymbol{Z}_{-1} \quad (T\bar{\boldsymbol{t}}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\boldsymbol{\iota}')\boldsymbol{W}\right]\begin{bmatrix} \hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}}_{OLS} \\ \hat{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}}_{OLS} \end{bmatrix}, \tag{3.B.4}$$

whereas for $\boldsymbol{B} \neq \boldsymbol{0}$ we will use the equivalent expression

$$\begin{aligned}\hat{\tau}_0 - \hat{\tau}_{0,OLS} =& \frac{1}{|\boldsymbol{D}'\boldsymbol{D}|}\left[(T\bar{\boldsymbol{t}}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\boldsymbol{\iota}')\boldsymbol{Z}_{-1}\boldsymbol{B} \quad (T\bar{\boldsymbol{t}}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\boldsymbol{\iota}')\boldsymbol{W} \quad (T\bar{\boldsymbol{t}}' - \boldsymbol{\iota}'\bar{\boldsymbol{t}}\boldsymbol{\iota}')\boldsymbol{Z}_{-1}\boldsymbol{A}_\perp\right] \\ &\times \begin{bmatrix} (\boldsymbol{A}'\boldsymbol{B})^{-1}\boldsymbol{A}'(\hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}}_{OLS}) \\ \hat{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}}_{OLS} \\ (\boldsymbol{B}_\perp'\boldsymbol{A}_\perp)^{-1}\boldsymbol{B}_\perp'(\hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}}_{OLS}) \end{bmatrix}.\end{aligned} \tag{3.B.5}$$

Then, for the case $\boldsymbol{\tau} = \boldsymbol{0}$ and $\boldsymbol{B} = \boldsymbol{0}$, (3.B.4) gives

$$\hat{\tau}_0 - \hat{\tau}_{0,OLS} = O(T^{-4})\left[O_p(T^{7/2}) \quad O_p(T^{5/2})\right]\begin{bmatrix} o_p(T^{-1}) \\ o_p(T^{-1/2}) \end{bmatrix} = o_p(T^{-3/2}).$$

For the case $\boldsymbol{\tau} = \boldsymbol{0}$ and $\boldsymbol{B} \neq \boldsymbol{0}$, (3.B.5) gives

$$\hat{\tau}_0 - \hat{\tau}_{0,OLS} = O(T^{-4})\left[O_p(T^3) \quad O_p(T^{5/2}) \quad O_p(T^{7/2})\right]\begin{bmatrix} o_p(T^{-1/2}) \\ o_p(T^{-1/2}) \\ o_p(T^{-1}) \end{bmatrix} = o_p(T^{-3/2}).$$

Next, assuming that $\tau \neq 0$ and $\beta = 0$, it follows from (3.B.4) that

$$\hat{\tau}_0 - \hat{\tau}_{0,OLS} = O(T^{-4}) \begin{bmatrix} O_p(T^4) & O_p(T^3) \end{bmatrix} \begin{bmatrix} o_p(T^{-1}) \\ o_p(T^{-1/2}) \end{bmatrix} = o_p(T^{-1}).$$

Alternatively, if $\boldsymbol{\tau} \neq \mathbf{0}$, $\boldsymbol{B} \neq \mathbf{0}$ and $\boldsymbol{B}'\boldsymbol{\tau} = \mathbf{0}$, then (3.B.5) gives

$$\hat{\tau}_0 - \hat{\tau}_{0,OLS} = O(T^{-4}) \begin{bmatrix} O_p(T^3) & O_p(T^3) & O_p(T^4) \end{bmatrix} \begin{bmatrix} o_p(T^{-1/2}) \\ o_p(T^{-1/2}) \\ o_p(T^{-1}) \end{bmatrix} = o_p(T^{-1}).$$

Finally, assume that $\boldsymbol{\tau} \neq \mathbf{0}$, and $\boldsymbol{B}'\boldsymbol{\tau} \neq \mathbf{0}$. Then, (3.B.5) gives

$$\hat{\tau}_0 - \hat{\tau}_{0,OLS} = O(T^{-4}) \begin{bmatrix} O_p(T^4) & O_p(T^3) & O_p(T^4) \end{bmatrix} \begin{bmatrix} o_p(T^{-1/2}) \\ o_p(T^{-1/2}) \\ o_p(T^{-1}) \end{bmatrix} = o_p(T^{-1/2}).$$

This completes the proof of Corollary 3.2. ■

3.B.2 Data Description

Variable	groups	Translation	Inclusion	Differenced
vakantiebaan	Job Search	holiday job	100%	N
Unemployment	Y	Unemployment	80%	Y
uwv vacatures	Job Search	uwv vacancies	78%	Y
werkloos	Social Security	unemployed	76%	Y
ww uitkering	Social Security	ww benefits	72%	Y
Ww	Social Security	Ww	69%	Y
nationale vacaturebank	RA	nationale vacaturebank	59%	Y
cv maken	Application training	CV write	57%	Y
indeed	RA	indeed	52%	Y
jobtrack	RA	jobtrack	52%	Y
motivatiebrief	Application training	motivation letter	52%	Y
sollicitatiebrief schrijven	Application training	write application letter	50%	Y
voorbeeld cv	Application training	example cv	48%	Y
tempo team	RA	tempo team	48%	Y
ontslagvergoeding	Social Security	severance pay	46%	Y
ww uitkering aanvragen	Social Security	request unemployment benefits	46%	Y
aanvragen uitkering	Social Security	request benefits	44%	N
interin	RA	interin	44%	Y
manpower	RA	manpower	44%	Y
randstad	General	randstad	44%	Y
werkzoekende	Social Security	job seeker	43%	Y
job	General	job	43%	Y
uwv	Social Security	uwv	43%	Y
werk.nl	Job Search	werk.nl	41%	Y
job vacancy	Job Search	job vacancy	41%	Y
uitkering	Social Security	benefits	41%	Y
ontslag	Social Security	resignation	41%	N
vacature	Job Search	vacancy	41%	Y
sollicitatiebrief voorbeeld	Application training	application letter example	39%	Y
sollicitatie	Application training	application	39%	Y
sollicitatiebrief	Application training	application letter	39%	Y
uitzendbureau	RA	employment agency	39%	Y
vakantiewerk	Job Search	holiday job	37%	N
tence	RA	tence	37%	Y

vacaturebank	Job Search	vacaturebank	37%	Y
sollicitatiegesprek	Application training	application interview	37%	N
tempo team uitzendbureau	RA	tempo team employment agency	35%	N
motivatiebrief voorbeeld	Application training	motivation letter example	35%	Y
bijstand	Social Security	social benefits	35%	Y
open sollicitatiebrief	Application training	open application letter	35%	Y
vrijwilligerswerk	General	volunteer work	35%	N
werk nl	Job Search	werk nl	35%	N
adecco	RA	adecco	33%	N
creyfs	RA	creyfs	33%	Y
randstad uitzendbureau	Job Search	randstad employment agency	33%	Y
cv maken voorbeeld	Application training	write CV example	31%	Y
werkbedrijf	Job Search	werkbedrijf	31%	Y
tempo-team	RA	tempo-team	31%	Y
werkloosheidsuitkering	Social Security	unemployment benefits	31%	N
tempo team vacatures	RA	tempo team vacancies	31%	Y
curriculum vitae voorbeeld	Application training	CV Example	31%	Y
cv	Application training	cv	31%	N
solliciteren	Application training	applying	31%	Y
indeed jobs	RA	indeed jobs	30%	Y
motivation letter	Application training	motivation letter	30%	N
resume example	Application training	resume example	28%	N
olympia uitzendbureau	RA	olympia employment agency	28%	Y
tempoteam	RA	tempoteam	28%	Y
randstad vacatures	Job Search	randstad vacancies	26%	Y
banen	General	jobs	26%	N
vrijwilliger	General	volunteer	26%	N
baan	General	job	26%	N
start uitzendbureau	RA	start employment agency	24%	Y
jobnet	RA	jobnet	24%	N
monsterboard	Job Search	monsterboard	24%	Y
baan zoeken	Job Search	job search	20%	N
functieomschrijving	General	job position description	20%	N
resume template	Application training	resume template	19%	N
omscholen	Application training	retraining	19%	Y
job interview	Application training	job interview	19%	N
werken bij	General	working at	19%	Y
vacatures	Job Search	vacancies	19%	Y
uwv uitkering	Social Security	uwv benefits	17%	Y
job description	General	job description	17%	Y
werk zoeken	General	job search	17%	Y
jobs	General	jobs	17%	Y
resumé	Application training	resume	15%	Y
bijscholen	Application training	retraining	15%	N
curriculum vitae template	Application training	CV Template	13%	N
curriculum vitae	Application training	CV	11%	Y
sollicitaties	Application training	applications	9%	Y
werkeloos	Social Security	unemployed	9%	N
werkloosheid	Social Security	unemployment	4%	N
resume	Application training	resume	2%	N
arbeidsbureau	RA	employment office	2%	N
uitzendbureaus	RA	employment agencies	2%	Y
werkloosheidswet	Social Security	unemployment law	0%	N

Chapter 4

High-Dimensional Single-Equation Cointegration Modelling

"Although modern computer technology helps us in so many respects, it also brings a new and urgent task to the statistician; that is, whether the classical limit theorems (i.e., those assuming a fixed dimension) are still valid for analyzing high dimensional data and how to remedy them if they are not."

- Bai and Silverstein (2010)

Abstract[†]

In this chapter, we extend the asymptotic theory for single-equation cointegration analysis from Chapter 3 to a high-dimensional framework. Sufficient conditions are derived under which the Single-equation Penalized Error Correction Selector attains simultaneous estimation and selection consistency. As the results strongly rely on the availability of suitable weights, we derive the consistency of the ridge estimator in our framework and illustrate how ridge may be used as an initial estimator for the construction of these weights. While consistency is attained, we demonstrate that the theoretically admissible growth rate of the integrated variables is slower than that of the stationary variables.

[†]This chapter is based on joint work with S. Smeekes.

4.1 Introduction

In this chapter, we extend the asymptotic theory for single-equation cointegration analysis to a high-dimensional framework. The theoretical properties of the Single-equation Penalized Error Correction Selector (SPECS) proposed in Chapter 3 are based on fixed-dimensional asymptotics. However, a key benefit of SPECS is that it enables estimation on high-dimensional datasets in which the cross-sectional dimension N is relatively large to the time series dimension T. In an attempt to obtain better asymptotic approximations in this setting, we demonstrate that SPECS maintains its attractive features, such as estimation and selection consistency, in an asymptotic framework in which the number of variables diverges. Moreover, we show that the fixed-dimensional results from Chapter 3 follow as a special case from the results presented in the current chapter.

The theoretical analysis of high-dimensional estimators frequently relies on the use of finite-sample bounds in which the dependence on the sample size and dimension are made explicit. In a stationary setting, the theory for L_1-penalized regression in high-dimensional settings is increasingly well-understood (e.g. Kock and Callot, 2015; Medeiros and Mendes, 2016). A popular method to gain insights into the theoretical properties of the lasso is to derive so-called oracle inequalities, which are sharp bounds on its prediction error and estimation error. To obtain these oracle inequalities, it is necessary to impose conditions that are strongly related to the eigenvalues of the scaled Gram matrix. Among the most used conditions are the *restricted eigenvalue condition* by Song and Bickel (2011) and the slightly more general *compatibility condition* that first appeared in Van de Geer (2007). An elaborate overview of these and related conditions, henceforth simply referred to as eigenvalue conditions, are provided in Van De Geer and Bühlmann (2009) and Bühlmann and Van De Geer (2011). While these eigenvalue conditions come in different levels of generality, they tend to be complicated to verify directly when the Gram matrix is random. A rather successful approach to circumvent this issue has been to assume an eigenvalue condition to hold on a simpler approximating matrix. It turns out that, when the approximation error vanishes sufficiently fast, the compatibility condition carries over to the scaled Gram matrix and the oracle inequalities can be derived in the usual fashion (e.g. Bühlmann and Van De Geer, 2011, Lemma 6.17). However, this approach does not extend easily to the non-stationary setting due to the lack of a simple non-random approximating matrix. Therefore, a key theoretical contribution in this chapter is related to the extension of eigenvalue conditions in the non-stationary setting.

The chapter proceeds as follows. In Section 4.2 we define the estimator and lay out our assumptions regarding the underlying DGP, the design and the required regularization. The main theoretical results for our estimator are presented in Section 4.3. In particular, the theoretical properties of SPECS and ridge are derived in Sections 4.3.1 and 4.3.2, respectively, and an illustrative example is provided in Section 4.3.3. Finally, we conclude in Section 4.4.

Notation

For any an N-dimensional vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^N x_i^p\right)^{1/p}$ denotes the ℓ_p-norm, while for any matrix $\boldsymbol{D}$ with N columns, $\|\boldsymbol{D}\|_p = \max_{\boldsymbol{x}\in R^N} \frac{\|\boldsymbol{D}x\|_p}{\|\boldsymbol{x}\|_p}$ is the corresponding induced norm. For an index set $S \subset \{1,\dots,N\}$, let $\boldsymbol{x}_S$ be the vector containing the elements of $\boldsymbol{x}$ corresponding to S. Similarly, for a matrix $\boldsymbol{D}$ with N rows, $\boldsymbol{D}_S$ is the submatrix containing the rows of $\boldsymbol{D}$ indexed by S. The orthogonal complement of $\boldsymbol{D}$ is denoted by $\boldsymbol{D}_\perp$, such that $\boldsymbol{D}_\perp'\boldsymbol{D} = \boldsymbol{0}$. When $\boldsymbol{D}$ is a square matrix, we denote its N ordered eigenvalues by $\lambda_1(\boldsymbol{D}) \geq \ldots \geq \lambda_N(\boldsymbol{D})$ and we use $\boldsymbol{D} \succ 0$ to denote that the matrix is positive definite. We use $\boldsymbol{\iota}_N$ to denote a vector of ones of length N and $\boldsymbol{I}_N$ to denote the N-dimensional identity matrix. We use $\xrightarrow{p}$ ($\xrightarrow{d}$) to denote convergence in probability (distribution) and $\overset{d}{=}$ denotes equivalence in distribution. Finally, we frequently make use of an arbitrary positive and finite constant K whose value may change throughout the paper, but is always independent of the time and cross-sectional dimensions.

4.2 Model, Estimator and Assumptions

The model that we consider here is analogous to that of Chapter 3. For convenience, we repeat the essential details. Assume that a researcher is interested in modelling a single variable of interest, say y_t, based on a N-dimensional time series $\boldsymbol{z}_t = (y_t, \boldsymbol{x}_t')$ that is observed for the periods $t = 1,\dots,T$. Furthermore, let $\boldsymbol{z}_t$ be described by the vector error correction model (VECM)

$$\Delta \boldsymbol{z}_t = \boldsymbol{A}\boldsymbol{B}'\boldsymbol{z}_{t-1} + \sum_{j=1}^{p} \boldsymbol{\Phi}_j \Delta \boldsymbol{z}_{t-j} + \boldsymbol{\epsilon}_t, \tag{4.2.1}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are $(N \times r)$-dimensional containing the adjustment rates and cointegrating vectors, respectively, and $\boldsymbol{\epsilon}_t = (\epsilon_{1,t}, \boldsymbol{\epsilon}_{2,t}')'$. Under suitable assumptions, defined in Section 4.2.3, the Granger Representation Theorem (e.g. Johansen, 1995a, p. 49), enables (4.2.1) to be written as a vector moving average (VMA) process of

the form

$$\boldsymbol{z}_t = \boldsymbol{C}\boldsymbol{s}_t + \boldsymbol{C}(L)\boldsymbol{\epsilon}_t + \boldsymbol{C}\boldsymbol{z}_0, \tag{4.2.2}$$

where $\boldsymbol{C} = \boldsymbol{B}_\perp \left(\boldsymbol{A}_\perp' \left(\boldsymbol{I}_N - \sum_{j=1}^{p} \boldsymbol{\Phi}_j \right) \boldsymbol{B}_\perp \right)^{-1} \boldsymbol{A}_\perp'$, $\boldsymbol{s}_t = \sum_{s=1}^{t} \boldsymbol{\epsilon}_s$, $\boldsymbol{C}(L)\boldsymbol{\epsilon}_t$ is a stationary linear process and $\boldsymbol{z}_0$ are initial values. Without loss of generality, we assume henceforth that $\boldsymbol{z}_0 = \boldsymbol{0}$. Typically, the finite-order VECM process is easier to estimate than the infinite order VMA process. Nonetheless, the number of parameters to estimate in (4.2.1) is at least $2Nr + N^2p$, such that the system quickly grows too large to accurately estimate based on traditional methods. Hence, from a computational perspective, an alternative lower-dimensional model formulation would be preferred.

4.2.1 Model

Utilizing that the modelling exercise focusses on a single variable of interest, the first form of dimension reduction that the researcher may wish to consider is to define a single-equation model for y_t. The importance in deriving a single-equation model for y_t is to ensure that the variables modelling the variation in y_t remain exogenous. This is accomplished by orthogonalizing the errors driving the single-equation model, say $\epsilon_{y,t}$, from the errors driving the marginal equation of the endogenous variables $\boldsymbol{x}_t$. Orthogonalization is achieved by decomposing $\epsilon_{1,t}$ into its best linear prediction based on $\boldsymbol{\epsilon}_{2,t}$ and the corresponding orthogonal prediction error. To this end, partition the covariance matrix of $\boldsymbol{\epsilon}_t$ as

$$\boldsymbol{\Sigma}_\epsilon = \begin{bmatrix} \sigma_{11} & \boldsymbol{\sigma}_{21}' \\ \boldsymbol{\sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

such that we obtain

$$\epsilon_{1,t} = (0, \boldsymbol{\sigma}_{21}' \boldsymbol{\Sigma}_{22}^{-1}) \boldsymbol{\epsilon}_t + \left(1, -\boldsymbol{\sigma}_{21}' \boldsymbol{\Sigma}_{22}^{-1}\right) \boldsymbol{\epsilon}_t = \hat{\epsilon}_{1,t} + \epsilon_{y,t}. \tag{4.2.3}$$

Define $\pi_0 = \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_{21}$. Then, writing out (4.2.3) in terms of the observable time series results in the single-equation model

$$\begin{aligned} \Delta \boldsymbol{y}_t &= (1, -\boldsymbol{\pi}_0') \left(\boldsymbol{A}\boldsymbol{B}' \boldsymbol{z}_{t-1} + \sum_{j=1}^{p} \boldsymbol{\Phi}_j' \Delta \boldsymbol{z}_{t-j} \right) + \boldsymbol{\pi}_0' \Delta \boldsymbol{x}_t + \epsilon_{y,t} \\ &= \boldsymbol{\delta}' \boldsymbol{z}_{t-1} + \boldsymbol{\pi}' \boldsymbol{w}_t + \epsilon_{y,t}, \end{aligned} \tag{4.2.4}$$

where $\boldsymbol{\delta}' = (1, -\boldsymbol{\pi}_0')\,\boldsymbol{A}\boldsymbol{B}'$ and $\boldsymbol{\pi} = (\boldsymbol{\pi}_0', \ldots, \boldsymbol{\pi}_p')'$ with $\boldsymbol{\pi}_j' = (1, -\boldsymbol{\pi}_0')\boldsymbol{\Phi}_j$ for $j = 1, \ldots, p$. Note that $\boldsymbol{\delta}$ is a vector of length N, whereas $\boldsymbol{\pi}$ is a vector of length $M = N(p+1) - 1$. Additionally, $\boldsymbol{w}_t = (\Delta\boldsymbol{x}_t', \Delta\boldsymbol{z}_{t-1}', \ldots, \Delta\boldsymbol{z}_{t-p}')'$ and $\epsilon_{y,t} = (1 - \boldsymbol{\pi}_0')\boldsymbol{\epsilon}_t$. Finally, we write the single-equation model in matrix notation as

$$\Delta\boldsymbol{y} = \boldsymbol{Z}_{-1}\delta + \boldsymbol{W}\pi + \boldsymbol{\epsilon}_y = \boldsymbol{V}\boldsymbol{\gamma} + \boldsymbol{\epsilon}_y, \tag{4.2.5}$$

where $\boldsymbol{V} = (\boldsymbol{Z}_{-1}, \boldsymbol{W})$, $\boldsymbol{Z}_{-1} = (\boldsymbol{z}_0, \ldots, \boldsymbol{z}_{T-1})'$, $\boldsymbol{W} = (\boldsymbol{w}_t, \ldots, \boldsymbol{w}_T)'$ and $\boldsymbol{\gamma} = (\boldsymbol{\delta}', \boldsymbol{\pi}')'$.

In deriving the theoretical properties of our estimator, it is useful to partition and rotate the data. Without loss of generality, we partition the data matrix as $\boldsymbol{V} = (\boldsymbol{V}_{S_\gamma}, \boldsymbol{V}_{S_\gamma^c})$, with $\boldsymbol{V}_{S_\gamma} = (\boldsymbol{Z}_{-1,S_\delta}, \boldsymbol{W}_{S_\pi})$ representing the time series carrying non-zero coefficients in the population single-equation model, henceforth referred as the set of relevant variables. In the presence of cointegration, it follows from (4.2.2) that the relevant lagged levels can be written as

$$\begin{aligned} \boldsymbol{z}_{S_\delta,t} &= \boldsymbol{C}_{S_\delta}\boldsymbol{s}_t + \boldsymbol{u}_{S_\delta,t}, \\ \boldsymbol{C}_{S_\delta} &= \boldsymbol{B}_{\perp,S_\delta}\left(\boldsymbol{A}_\perp'\left(\boldsymbol{I}_N - \sum_{j=1}^{p}\boldsymbol{\Phi}_j\right)\boldsymbol{B}_\perp\right)^{-1}\boldsymbol{A}_\perp' \end{aligned} \tag{4.2.6}$$

where $\boldsymbol{B}_{\perp,S_\delta}$ is an $(|S_\delta| \times (N-r))$-dimensional matrix containing the rows of $\boldsymbol{B}_\perp$ indexed by S_δ. The left null space of $\boldsymbol{B}_{\perp,S_\delta}$, defined as

$$\boldsymbol{B}^* = \left\{\boldsymbol{x} \in \mathbb{R}^{|S_\delta|} \mid \boldsymbol{B}_{\perp,S_\delta}'\boldsymbol{x} = \boldsymbol{0}\right\},$$

contains the linear combinations that convert $\boldsymbol{z}_{S_\delta,t}$ to a stationary process. Accordingly, we also refer to this null space as the cointegrating space of $\boldsymbol{z}_{S_\delta,t}$. By construction, $\boldsymbol{\delta}_{S_\delta} \in \boldsymbol{B}^*$, such that this cointegrating space is non-empty whenever $\boldsymbol{\delta} \neq \boldsymbol{0}$. In this case, we define $\boldsymbol{B}_{S_\delta}$ as a $(|S_\delta| \times r^*)$-dimensional basis matrix of $\boldsymbol{B}^*$, with $r^* \leq |S_\delta|$ representing the dimension of the cointegrating space.[1] Similarly, we define $\boldsymbol{B}_{S_\delta,\perp}$ as a basis matrix of the left null-space of $\boldsymbol{B}_{S_\delta}$, i.e. a $(|S_\delta| \times (|S_\delta| - r^*))$-dimensional matrix of full column rank with the property that $\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{B}_{S_\delta} = \boldsymbol{0}$. Then, we are able to define a $\boldsymbol{Q}$-transformation that decomposes the reduced system into a stationary

[1] The matrix $\boldsymbol{B}_{S_\delta}$ is not uniquely defined. However, in most instances, including those contained in the current chapter, identification of the span of $\boldsymbol{B}_{S_\delta}$ is sufficient.

and non-stationary contribution. Define,

$$\begin{aligned} \boldsymbol{Q} &= \begin{bmatrix} \boldsymbol{B}'_{S_\delta} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{|S_\pi|} \\ \boldsymbol{B}'_{S_\delta,\perp} & \mathbf{0} \end{bmatrix}, \text{ with} \\ \boldsymbol{Q}^{-1} &= \begin{bmatrix} \boldsymbol{B}_{S_\delta}\left(\boldsymbol{B}'_{S_\delta}\boldsymbol{B}_{S_\delta}\right)^{-1} & \mathbf{0} & \boldsymbol{B}_{S_\delta,\perp}\left(\boldsymbol{B}'_{S_\delta,\perp}\boldsymbol{B}_{S_\delta,\perp}\right)^{-1} \\ \mathbf{0} & \boldsymbol{I}_{|S_\pi|} & 0 \end{bmatrix}. \end{aligned} \tag{4.2.7}$$

Post-multiplication of the data matrix by Q gives

$$\boldsymbol{V}_{S_\gamma}\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Z}_{-1,S_\delta}\boldsymbol{B}_{S_\delta} & \boldsymbol{W}_{S_\pi} & \boldsymbol{Z}_{-1,S_\delta}\boldsymbol{B}_{S_\delta,\perp} \end{bmatrix} \tag{4.2.8}$$

which we refer to as the $\boldsymbol{Q}$-transformed version of $\boldsymbol{V}_{S_\gamma}$.

Remark 4.1. In an attempt to simplify the proofs in this chapter, we proceed under the assumption that $\boldsymbol{\delta} \neq \mathbf{0}$, i.e. $|S_\delta| \geq 1$. We believe that this assumption does not harm the generality of our results, as in a high-dimensional non-stationary time series setting it seems unrealistic that no cointegration appears in the single-equation model. In Chapter 3, however, we do allow for the case $\boldsymbol{\delta} = \mathbf{0}$, by defining a separate rotation and scaling matrix. A similar strategy is possible, though not pursued, in the current setting.

4.2.2 Estimator

Despite the dimension reduction obtained from moving towards a single-equation representation, regularization remains a necessity in high dimensions. The single-equation model (4.2.4) contains a total of $N(p+2)-1$ parameters, compared to the $2Nr + N^2p$ parameters in the full-system VECM in (4.2.1), resulting in a substantial reduction in dimensionality. However, the dimension may still grow large when either: (i) the number of potentially relevant variables is large or (ii) when the number of lagged differences required to appropriately model the short-run dynamics is large. Therefore, similar to Chapter 3, we consider the use of a shrinkage estimator for (4.2.4) that enables estimation in high-dimensions. In the previous chapter, the proposed version of SPECS incorporates a combination of both an L_1-penalty on the individual coefficients and an L_2-penalty on $\boldsymbol{\delta}$. While the latter penalty is intuitively motivated to enforce sparsity in the absence of cointegration, i.e. $\boldsymbol{\delta} = 0$, the results in Chapter 3 demonstrate that the L_2-penalty makes little difference. Moreover, the theory in this chapter is derived under the assumption that $\boldsymbol{\delta} \neq 0$. Therefore, we proceed without the additional L_2-penalty, simplifying the theoretical derivations and enabling

emphasis on the key issues in the high-dimensional analysis of non-stationary time series.

The estimator, for convenience still referred to as SPECS throughout this chapter, is defined as the minimizer of the following objective function:

$$G_T(\boldsymbol{\gamma}) = \|\Delta \boldsymbol{y} - \boldsymbol{V}\boldsymbol{\gamma}\|_2^2 + \lambda_T \sum_{i=1}^{N+M} \omega_i |\gamma_i| . \tag{4.2.9}$$

Indeed, this is the adaptive lasso, as defined in Zou (2006), applied to the conditional error correction model. The weights ω_i in (4.2.9) are typically derived from an initial estimation procedure, although to maintain generality we do not propose a particular construction at this stage. As demonstrated by Zou (2006), under certain assumptions on the weights, the adaptive lasso attains simultaneous selection and estimation consistency, without the necessity for the rather stringent irrepresentability condition in Zhao and Yu (2006). In pursuit of similar theoretical properties for SPECS, we define the appropriate assumptions on the weights, among others, in the following section.

4.2.3 Assumptions

In this section we define and discuss the assumptions required for the main results in this chapter. First, the following assumptions are imposed on the innovations.

Assumption 4.1. The sequence of innovations $\{\boldsymbol{\epsilon}_t\}_{t\geq 1}$ is an N-dimensional martingale difference sequence (m.d.s.) with $\mathbb{E}(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t') = \boldsymbol{\Sigma}_\epsilon$. Furthermore, we assume that

1. there exists a $m > 2$, such that $\max_{1\leq i\leq N, 1\leq t\leq T} \mathbb{E}|\epsilon_{i,t}|^{2m} \leq K_m$, and
2. there exist $\phi_{\min}, \phi_{\max} > 0$, such that $\phi_{\min} \leq \lambda_{\min}(\boldsymbol{\Sigma}_\epsilon) < \lambda_{\max}(\boldsymbol{\Sigma}_\epsilon) \leq \phi_{\max}$.

The first part of Assumption 4.1 is required for the application of a high-dimensional law of large numbers in Lemma 4.4 in the Appendix. In the second part, the lower bound on the minimum eigenvalue of the covariance matrix is necessary to ensure that the eigenvalues of the sample covariance matrix are bounded away from zero, whereas the upper bound on the maximum eigenvalue is helpful in showing convergence of certain sample covariance matrices.

Next, we require that the VECM model admits the vector moving average (VMA) representation displayed in (4.2.2). By the Granger Representation Theorem, the following assumptions are sufficient.

Assumption 4.2. Define $\boldsymbol{A}(z) := (1-z)\boldsymbol{I}_N - \boldsymbol{A}\boldsymbol{B}'z - \sum_{j=1}^{p}\boldsymbol{\Phi}_j(1-z)z^j$.

(i) The determinantal equation $|\boldsymbol{A}(z)|$ has all roots on or outside the unit circle.

(ii) $\boldsymbol{A}$ and $\boldsymbol{B}$ are $N \times r$ matrices with $1 \leq r \leq N$ and $\text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{B}) = r$.

(iii) The $((N-r)\times(N-r))$ matrix $\boldsymbol{A}'_\perp\left(I_N - \sum_{j=1}^{p}\boldsymbol{\Phi}_j\right)\boldsymbol{B}_\perp$ is invertible.

The existence of a VMA representation alone is not sufficient for the convergence of our high-dimensional sample covariance matrices. A further restriction on the dependency over time is required, in the form of the following assumption.

Assumption 4.3. There exists a finite K such that the matrix $\boldsymbol{C}$ in (4.2.2) satisfies $\|\boldsymbol{C}\|_\infty \leq K$. In addition, the matrix lag polynomial $\boldsymbol{C}(L)$ is given by $\boldsymbol{C}(z) = \sum_{l=0}^{\infty}\boldsymbol{C}_l z^l$ and satisfies $\sum_{l=0}^{\infty} l\,\|\boldsymbol{C}_l\|_\infty \leq K$.

Assumption 4.3 is particularly useful in ensuring norm-summability of the coefficients in the Beveridge-Nelson decomposition. More precisely, we may decompose $\boldsymbol{C}(z) = \boldsymbol{C}(1) + (1-z)\boldsymbol{C}^*(z)$, where $\boldsymbol{C}^*(z) = \sum_{l=0}^{\infty}\boldsymbol{C}_l^*$ with $\boldsymbol{C}_l^* = -\sum_{k=l+1}^{\infty}\boldsymbol{C}_l$. It follows that,

$$\sum_{l=0}^{\infty}\|\boldsymbol{C}_l^*\|_\infty = \sum_{l=0}^{\infty}\left\|\sum_{k=l+1}^{\infty}\boldsymbol{C}_l\right\|_\infty \leq \sum_{l=0}^{\infty}\sum_{k=l+1}^{\infty}\|\boldsymbol{C}_l\|_\infty = \sum_{l=1}^{\infty} l\,\|\boldsymbol{C}_l\|_\infty < \infty.$$

This property is used to bound several quantities of interest in the proofs of our theoretical results.

The ability of our estimation procedure to consistently select and estimate the coefficients of the relevant variables hinges on the behaviour of the eigenvalues of the sample covariance matrices. Under Assumptions 4.1-4.3, it is possible to ensure eigenvalue conditions on the sample covariance matrices, by imposing them on simpler approximating matrices. Accordingly, we make the following assumption.

Assumption 4.4. Define $\boldsymbol{v}_{1,t} = \left(\boldsymbol{z}'_{S_\delta,t}\boldsymbol{B}_{S_\delta}, \boldsymbol{w}'_{S_\pi,t}\right)'$, $\boldsymbol{v}_{2,t} = \boldsymbol{B}'_{S_\delta,\perp}\boldsymbol{z}_{S_\delta,t}$, $s_\pi = |S_\pi| + r^*$ and $s_\delta = |S_\delta| - r^*$. Furthermore, let $\hat{\boldsymbol{\Sigma}}_{11} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{v}_{1,t}\boldsymbol{v}'_{1,t}$ and $\hat{\boldsymbol{\Sigma}}_{22} = \frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{v}_{2,t}\boldsymbol{v}'_{2,t}$. Then, we assume that

1. There exists a constant $\phi > 0$, such that

$$\inf_{\boldsymbol{x}\in R^{s_\pi}}\frac{\boldsymbol{x}'\hat{\boldsymbol{\Sigma}}_{11}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} \geq \phi. \tag{4.2.10}$$

2. Similarly, it holds that,

$$\inf_{\boldsymbol{x}\in R^{s_\delta}}\frac{\boldsymbol{x}'\hat{\boldsymbol{\Sigma}}_{22}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} \geq \phi, \tag{4.2.11}$$

with probability converging to 1 as $T, N, s_\pi, s_\delta \to \infty$.

The first part of Assumption 4.4 applies to stationary data and is known to hold when the minimum eigenvalue of the corresponding population covariance matrix is bounded away from zero (e.g. Medeiros and Mendes, 2016, Section B.2). The second part, however, applies to integrated variables and requires arguments that are unique to the non-stationary setting. In particular, we note the necessity of applying a scaling by $\frac{s_\delta}{T^2}$, rather than the usual $\frac{1}{T^2}$ one may expect from the fixed-dimensional literature, cf. Remark 4.2. In Appendix 4.A.3, we show several cases under which Assumption 4.4 is satisfied.

Remark 4.2. As an illustration of the problems with adopting the usual scaling by T^{-2}, consider the simple example of an s-dimensional white noise sequence $\boldsymbol{u}_t \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_s)$ and define $\boldsymbol{h}_t = \sum_{j=1}^{t} \boldsymbol{u}_j$. Then, in Lemma 4.5 in Appendix 4.A.3 we show that $\mathbb{P}\left(\lambda_{\min}\left(\frac{1}{T^2}\sum_{t=1}^{T} \boldsymbol{h}_t\boldsymbol{h}_t'\right) > \phi\right) \to 0$, as $s, T \to \infty$, regardless of their relative rates. Hence, even in this simple case we cannot assume that the minimum eigenvalue is bounded away from zero if we stick to the T^{-2} scaling.

Remark 4.3. There are several noteworthy instances in which $\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{22}\right)$ is bounded away from zero with arbitrarily high probability without the need for Assumption 4.4. In particular, assume that the dimension of the orthogonal complement of the cointegrating space in the subset of relevant non-stationary variables converges to a finite constant, i.e. $s_\delta(T) \to K$ as $T \to \infty$. Then, based on the functional central limit theorem

$$\hat{\boldsymbol{\Sigma}}_{22} \xrightarrow{d} K\boldsymbol{B}'_{S_\delta,\perp}\boldsymbol{C}_{S_\delta}\left(\int_0^1 \boldsymbol{B}(r)\boldsymbol{B}'(r)dr\right)\boldsymbol{C}'_{S_\delta}\boldsymbol{B}_{S_\delta,\perp} \overset{d}{=} \int_0^1 \boldsymbol{B}^*(r)\boldsymbol{B}^{*\prime}(r)dr,$$

where $\boldsymbol{B}^*(r)$ is an s_δ-dimensional Brownian Motion with $\mathbb{E}\left(\boldsymbol{B}^*(r)\boldsymbol{B}^{*\prime}(r)\right) = rK^2\boldsymbol{B}'_{S_\delta,\perp}\boldsymbol{C}_{S_\delta}\boldsymbol{\Sigma}_\epsilon\boldsymbol{C}'_{S_\delta}\boldsymbol{B}_{S_\delta,\perp}$. By Lemma A.2 in Phillips and Hansen (1990), it follows that $\int_0^1 \boldsymbol{B}^*(r)\boldsymbol{B}^{*\prime}(r)dr$ is positive-definite almost surely. Then, by continuity of the eigenvalue, we may choose $\phi(\epsilon) > 0$ such that

$$\mathbb{P}\left(\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{22}\right) \geq \phi(\epsilon)\right) \to \mathbb{P}\left(\lambda_{\min}\left(\int_0^1 \boldsymbol{B}^*(r)\boldsymbol{B}^{*\prime}(r)dr\right) \geq \phi(\epsilon)\right) \leq \epsilon,$$

for any $\epsilon > 0$. A straightforward case in which s_δ remains finite is to simply assume that the number of relevant integrated variables, i.e. $|S_\delta|$, stays finite. However, a more general example occurs when the dimension of the cointegrating space of $\boldsymbol{z}_{S_\delta,t}$ diverges at the rate $|S_\delta|$. This occurs in the case of a non-stationary factor model with idiosyncratic components, as proposed by Banerjee et al. (2014a). Further illustrations are provided in Remark 4.8.

Finally, to ensure simultaneous recovery of the correct sparsity patterns and consistent estimation of the non-zero coefficients, we impose a set of conditions on the tuning parameter λ_T and the weights $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_{N+M})'$.

Assumption 4.5. Assume that the following claims hold.

1. The smallest population coefficient is allowed to decrease to zero, as long as

$$\frac{|\gamma_{\min}|\sqrt{T}}{(s_\delta \vee \sqrt{s_\pi})} \to \infty.$$

2. The penalty parameter grows sufficiently slow, such that

$$\frac{\lambda_T \left(\sqrt{s_\delta} \vee \sqrt{s_\pi}\right)}{\left(s_\delta \vee \sqrt{s_\pi}\right) T^{1/2-\xi}} \to 0,$$

   for some constant $\xi > 0$.

3. The weights corresponding to the relevant variables satisfy

$$\omega_{S_\gamma,\max} \leq T^\xi,$$

   with probability approaching one.

4. The weights corresponding to the irrelevant variables and the penalty parameter grow sufficiently fast:

$$\frac{\omega_{S_\delta^c,\min}}{\left(\sqrt{s_\delta} \vee \sqrt{s_\pi}\right) T^{1/2+\xi}\sqrt{N}} \to \infty, \qquad \frac{\lambda_T \omega_{S_\delta^c,\min}}{(s_\delta \vee \sqrt{s_\pi})T\sqrt{N}} \to \infty,$$
$$\frac{\omega_{S_\pi^c,\min}}{\left(\sqrt{s_\delta} \vee \sqrt{s_\pi}\right) T^{\xi}\sqrt{M}} \to \infty, \qquad \frac{\lambda_T \omega_{S_\pi^c,\min}}{(s_\delta \vee \sqrt{s_\pi})\sqrt{TM}} \to \infty.$$

The first part of Assumption 4.5 determines the fastest rate at which the population coefficient is allowed to decrease, as a function of the growth rates of s_δ and s_π. Intuitively, the faster the number of relevant variables diverges, the slower the minimum coefficient may go to zero to ensure identification of small non-zero coefficients. The maximum rates of s_δ and s_π are specified in Theorem 4.1. The second part puts an upper bound on the admissible growth rate of the penalty. Exceeding this bound result in an excess of shrinkage bias that impedes estimation consistency. For the same reason, the third part requires that the weights of the relevant variables do not grow too fast. Finally, part four states that the penalty parameter and the weights of the irrelevant variables grow sufficiently fast in order to guarantee that

irrelevant variables are removed from the model with probability converging to one. The required minimum growth rate of the penalty parameter is inversely related to the growth rate of the weights of the irrelevant variables; faster diverging weights require less penalization to identify irrelevant variables.

4.3 Theoretical Results

In this section we derive the asymptotic properties of SPECS, describe the construction of the weights and provide illustrative examples in which we implement SPECS and obtain specific rates of convergence.

4.3.1 Main Theorems

The first result that we pursue is the selection consistency of our estimator, described in the following theorem.

Theorem 4.1. *Assume that* $\frac{s_\delta}{T^{1/4}} \to 0$ *and* $\frac{s_\pi}{\sqrt{T}} \to 0$. *Then, under Assumptions 4.1-4.5, it holds that*

$$\mathbb{P}\left(\operatorname{sign}\left(\hat{\boldsymbol{\gamma}}\right) = \operatorname{sign}\left(\boldsymbol{\gamma}\right)\right) \to 1,$$

as $T, N, p, s_\delta, s_\pi \to \infty$.

Theorem 4.1 states that the identified set of relevant variables corresponds to the true set with probability converging to one. This result provides an asymptotic justification for implementing SPECS as a high-dimensional variable selection device. Since the set of variables included is strictly smaller than the time series dimension, it is possible to apply a traditional consistent estimator to the selected set of variables (e.g. Belloni and Chernozhukov, 2013). However, ideally SPECS would contain desirable properties that omit the need of a second estimation procedure. For this reason, we establish the simultaneous consistency of the estimated coefficients in the following theorem.

Theorem 4.2. *Let* $\boldsymbol{S}_T = \operatorname{diag}\left(\sqrt{T}\boldsymbol{I}_{s_\pi}, \frac{T}{\sqrt{s_\delta}}\boldsymbol{I}_{s_\delta}\right)$ *and* $\boldsymbol{Q}$ *as defined in* (4.2.7). *Under the same assumption as in Theorem 4.1, it holds that*

$$\left\|\boldsymbol{S}_T\boldsymbol{Q}'^{-1}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma}\right)\right\|_2 = O_p\left(s_\delta \vee \sqrt{s_\pi}\right). \tag{4.3.1}$$

In the case where $s_\delta, s_\pi \leq K$, for some constant K that is independent of T, Theorem 4.2 is equivalent to Theorem 3.1. However, the current setting does not require that $N, M \leq K$, i.e. the number of irrelevant variables are allowed to diverge

without affecting the convergence rate of the estimator. Consequently, the current results nest those of Chapter 3 as a special case, while allowing for a more general asymptotic framework.

Remark 4.4. By the assumption on s_δ, it holds that $\frac{T}{\sqrt{s_\delta}} \geq \sqrt{T}$ for sufficiently large T, such that

$$\left\| \boldsymbol{S}_T \boldsymbol{Q}'^{-1} \left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma} \right) \right\|_2 \geq \sqrt{T} \left\| \boldsymbol{Q}'^{-1} \left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma} \right) \right\|_2 .$$

Moreover, since the basis matrices $\boldsymbol{B}_{S_\delta}$ and $\boldsymbol{B}_{S_\delta,\perp}$ are not uniquely defined, we may impose a normalization such that $\|\boldsymbol{Q}\|_2 \leq 1$. Then,

$$\begin{aligned} \left\| \hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma} \right\|_2 &= \left\| \boldsymbol{Q}' \boldsymbol{Q}'^{-1} \left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma} \right) \right\|_2 \\ &\leq \|\boldsymbol{Q}\|_2 \left\| \boldsymbol{Q}'^{-1} \left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma} \right) \right\|_2 \leq \left\| \boldsymbol{Q}'^{-1} \left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma} \right) \right\|_2 , \end{aligned}$$

such that

$$\left\| \boldsymbol{S}_T \boldsymbol{Q}'^{-1} \left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma} \right) \right\|_2 \geq \sqrt{T} \left\| \hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma} \right\|_2 .$$

Thus, it follows from Theorem 4.2 that

$$\left\| \hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma} \right\|_2 = O_p \left(\frac{\left(s_\delta \vee \sqrt{s_\pi} \right)}{\sqrt{T}} \right) .$$

It is immediate that SPECS attains $\sqrt{T}$-consistency when $s_\delta, s_\pi \leq K$, i.e. the convergence rate in a fixed-dimensional framework is equivalent to that of the OLS estimator.

4.3.2 Initial Estimates

In this section, we propose the use of the ridge estimator for the construction of initial weights, and derive its consistency under a further restriction of the asymptotic framework. Recall that the ridge estimator is defined as the minimizer of the following objective function:

$$G_R(\boldsymbol{\gamma}) := \|\Delta \boldsymbol{y} - \boldsymbol{V}\boldsymbol{\gamma}\|_2^2 + \lambda_R \|\boldsymbol{\gamma}\|_2^2 . \tag{4.3.2}$$

The properties of the ridge estimator are well-studied in the stationary setting (e.g Hastie et al., 2008, Section 3.4.1). However, to the best of our knowledge, no explicit results are available in the high-dimensional non-stationary case considered here.

A crucial assumption for the main theorems to hold, is the availability of suitable weights. Intuitively, the weights corresponding to the relevant variables should not

increase too fast to maintain estimation consistency, whereas those corresponding to the irrelevant variables should increase sufficiently fast to ensure selection consistency. To construct the weights, one commonly relies on initial estimates, say $\hat{\boldsymbol{\gamma}}_I$, obtained from a consistent estimator (e.g. Zou, 2006; Huang et al., 2008; Kock, 2016; Smeekes and Wijler, 2018a). Similar to the construction in Chapters 2 and 3, we define the weights as $\omega_i = \frac{1}{|\hat{\boldsymbol{\gamma}}_{I,i}|^k}$. This specification allows for substantial flexibility in the regulation of the divergence rate of weights corresponding to irrelevant variables. To illustrate, assume that $\hat{\gamma}_{I,i} = \gamma_i + O_p\left(T^{-a}\right)$ for all i. Then, it is clear that $\omega_i = O_p(1)$ when $\gamma_i \neq 0$ and $\omega_i = O_p\left(T^{ka}\right)$ when $\gamma_i = 0$. Therefore, larger values of k will increase the rate at which the weights corresponding to the irrelevant variables diverge. Based on this principle, the availability of a consistent initial estimator allows us to construct weights that satisfy the conditions in Assumption 4.5.

Remark 4.5. While the idea of adjusting divergence rates through imposing varying values of k seems theoretically attractive, large values of k result in substantial amplification of finite-sample estimation error. As a result, the finite-sample performance of the lasso becomes unstable for large k, such that in practice one may want to set the value for k as low as theoretically admissible.

Demonstrating the availability of a consistent initial estimator in the high-dimensional setting considered here requires the development of novel theoretical results. In an application where N is small relative to T, initial OLS estimates can be used and when N is close to or exceeding T, initial ridge estimates are a sensible choice. However, the properties of these estimators are unknown in the high-dimensional framework considered here. As an alternative, Huang et al. (2008) propose the use of marginal regression under a so-called 'partial orthogonality condition', which puts a restriction on the degree of correlation between the relevant and irrelevant variables. Unfortunately, in the non-stationary setting, such an assumption is unlikely to hold as a result of the correlation induced by common stochastic trends. A different promising option is to rely on initial (unweighted) lasso estimates. To validate this approach, however, the consistency and convergence rate of the lasso estimator needs to be derived in the current framework. The fastest way to derive consistency of the lasso, is through the use of a compatibility condition as in Bühlmann and Van De Geer (2011, Ch. 6). However, in addition to the difficulty of showing the theoretical validity of a compatibility condition in the non-stationary setting considered here, the use of a compatibility condition is further complicated by the fact that the stochastic trends asymptotically dominate the variation. More specifically, in order to attain a non-singular limit matrix, a rotation similar to $\boldsymbol{Q}$ is required that separates the stationary and non-stationary components in the full dataset. Accordingly, the standard com-

patibility condition needs to be adjusted in a non-trivial manner to account for such a rotation. Consequently, we prefer to rely on the ridge estimator, while postponing the use of initial lasso estimates based on a compatibility condition to future research.

In order to derive consistency of the ridge estimator, we extend the minimum eigenvalue bound in Assumption 4.4 as follows.

Assumption 4.6. Let $N_\delta = N - r$, $M_\pi = M + r$, $\boldsymbol{v}_{R1,t} = (\boldsymbol{z}_t'\boldsymbol{B}, \boldsymbol{w}_t')'$ and $\boldsymbol{v}_{R2,t} = \boldsymbol{B}_\perp'\boldsymbol{z}_t$. Furthermore, define $\hat{\boldsymbol{\Sigma}}_{R,11} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{v}_{R1,t}\boldsymbol{v}_{R1,t}'$ and $\hat{\boldsymbol{\Sigma}}_{R,22} = \frac{N_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{v}_{R2,t}\boldsymbol{v}_{R2,t}'$. Then, we assume that

1. There exists a constant $\phi_R > 0$, such that

$$\inf_{\boldsymbol{x}\in R^{M_\pi}} \frac{\boldsymbol{x}'\hat{\boldsymbol{\Sigma}}_{R,11}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} \geq \phi_R. \tag{4.3.3}$$

2. Similarly, it holds that,

$$\inf_{\boldsymbol{x}\in R^{N_\delta}} \frac{\boldsymbol{x}'\hat{\boldsymbol{\Sigma}}_{R,22}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} \geq \phi_R, \tag{4.3.4}$$

with probability converging to 1 as $T, N, p \to \infty$.

After controlling the minimum eigenvalue of the covariance matrices, we are able to derive the rate of convergence of the ridge estimator under a further restriction on the growth rates of N, M. The consistency of the ridge estimator is described in the following theorem.

Theorem 4.3. *Define the scaling and rotation matrices* $\boldsymbol{S}_R = diag\left(\sqrt{T}\boldsymbol{I}_{M_\pi}, \frac{T}{N_\delta}\boldsymbol{I}_{N_\delta}\right)$ *and*

$$\boldsymbol{Q}_R = \begin{bmatrix} (\boldsymbol{B}'\boldsymbol{B})^{-1/2}\boldsymbol{B}' & 0 \\ 0 & \boldsymbol{I}_M \\ (\boldsymbol{B}_\perp'\boldsymbol{B}_\perp)^{-1/2}\boldsymbol{B}_\perp' & 0 \end{bmatrix}.$$

Assume that $\frac{N_\delta}{T^{1/4}} \to 0$, $\frac{M_\pi}{\sqrt{T}} \to 0$, *and* $\lambda_R \leq \frac{\left(N_\delta \vee \sqrt{M_\pi}\right)\sqrt{T}}{\left(\sqrt{s_\delta} \vee \sqrt{s_\pi}\right)}$. *Then, under Assumptions 4.1-4.3 and 4.6, it holds that*

$$\left\|\boldsymbol{S}_R\boldsymbol{Q}_R'^{-1}\left(\hat{\gamma}_R - \boldsymbol{\gamma}\right)\right\|_2 = O_p\left(N_\delta \vee \sqrt{M_\pi}\right). \tag{4.3.5}$$

The attentive reader may note that the admissible growth rates of N_δ, M_π on Theorem 4.3 are the same as those initially assumed on the subsets of relevant variables,

i.e. s_δ, s_π, in Theorem 4.1. Ideally, we would like to allow for faster rates of divergence for the set of the irrelevant variables. Unfortunately, without the availability of a compatibility condition that could justify the plain lasso as an initial estimator, this restriction seems unavoidable. Nonetheless, several interesting and practically relevant settings exist where the generality provided by the asymptotic framework of Theorem 4.3 is sufficient, as is illustrated in the following section.

Remark 4.6. Similar to Remark 4.4, it follows from Theorem 4.3 that

$$\|\hat{\boldsymbol{\gamma}}_R - \boldsymbol{\gamma}\|_2 = O_p\left(\frac{\left(N_\delta \vee \sqrt{M_\pi}\right)}{\sqrt{T}}\right).$$

Based on the assumption that $\frac{N_\delta}{T^{1/4}} \to 0$ and $\frac{M_\pi}{\sqrt{T}} \to 0$ in Theorem 4.3, it follows directly that $\|\hat{\boldsymbol{\gamma}}_R - \boldsymbol{\gamma}\|_2 = o_p(1)$.

Remark 4.7. Theorem 4.3 imposes no minimum growth rate of the penalty term λ_R in (4.3.2). Therefore, in the case where $M + N < T$, the choice $\lambda_R = 0$ is both theoretically admissible and computationally feasible, such that consistency of the OLS estimator follows as a by-product of our result.

4.3.3 An Illustrative Example

We conclude our theoretical results by providing an illustrative example of a general DGP in an asymptotic framework that complies with the assumptions in Theorems 4.1-4.3. The required weights are explicitly constructed by means of an initial ridge estimator and particular rates of convergence of both the initial and final estimator are provided.

Assume that the researcher observes the N-dimensional time series

$$\boldsymbol{z}_t = (\boldsymbol{z}_{1,t}', \boldsymbol{z}_{2,t}')' = (y_t, \boldsymbol{x}_t')',$$

from time $t = 1, \ldots, T$, where $\boldsymbol{z}_{1,t} = (y_t, \boldsymbol{x}_{1,t}')'$ is an N_1-dimensional time series and $\boldsymbol{z}_{2,t}$ is an N_2-dimensional time series. Moreover,

$$\begin{aligned}\begin{bmatrix}\Delta \boldsymbol{z}_{1,t} \\ \Delta \boldsymbol{z}_{2,t}\end{bmatrix} &= \begin{bmatrix}\boldsymbol{\Pi}_{11} & \boldsymbol{\Pi}_{12} \\ \boldsymbol{\Pi}_{21} & \boldsymbol{\Pi}_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{z}_{1,t-1} \\ \boldsymbol{z}_{2,t-1}\end{bmatrix} + \sum_{j=1}^{p}\begin{bmatrix}\boldsymbol{\Phi}_{j,11} & \boldsymbol{\Phi}_{j,12} \\ \boldsymbol{\Phi}_{j,21} & \boldsymbol{\Phi}_{j,22}\end{bmatrix}\begin{bmatrix}\Delta \boldsymbol{z}_{1,t-j} \\ \Delta \boldsymbol{z}_{2,t-j}\end{bmatrix} + \begin{bmatrix}\boldsymbol{\epsilon}_{1,t} \\ \boldsymbol{\epsilon}_{2,t}\end{bmatrix} \\ &= \boldsymbol{\Pi}\boldsymbol{z}_{t-1} + \sum_{j=1}^{p}\boldsymbol{\Phi}_j \Delta \boldsymbol{z}_{t-j} + \boldsymbol{\epsilon}_t,\end{aligned} \tag{4.3.6}$$

where $\boldsymbol{\Pi}_{11} = \boldsymbol{A}_1\boldsymbol{B}_1'$ is an $(N_1 \times N_1)$-dimensional matrix with $\operatorname{rank}(\boldsymbol{\Pi}_{11}) = r_1$. In

addition, assume that $\boldsymbol{\Sigma}_\epsilon = \mathbb{E}\left(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t'\right)$ satisfies Assumption 4.1 and can be decomposed as

$$\begin{aligned} \boldsymbol{\Sigma}_\epsilon &= \begin{bmatrix} \boldsymbol{\Sigma}_{\epsilon,11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\epsilon,22} \end{bmatrix}, \text{ with } \boldsymbol{\Sigma}_{\epsilon,11} = \begin{bmatrix} \sigma_{1,11} & \boldsymbol{\sigma}_{1,21}' \\ \boldsymbol{\sigma}_{1,21} & \boldsymbol{\Sigma}_{1,22} \end{bmatrix} \text{ and} \\ \boldsymbol{\Sigma}_{\epsilon,22} &= \begin{bmatrix} \sigma_{2,11} & \boldsymbol{\sigma}_{2,21}' \\ \boldsymbol{\sigma}_{2,21} & \boldsymbol{\Sigma}_{2,22} \end{bmatrix}. \end{aligned} \tag{4.3.7}$$

Then, the quantities appearing in the construction of the single-equation model in (4.2.4) take on the form

$$\begin{aligned} \boldsymbol{\pi}_0 &= \begin{bmatrix} \boldsymbol{\Sigma}_{1,22}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\epsilon,22}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\sigma}_{1,21} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\pi}_{0,1} \\ \mathbf{0} \end{bmatrix}, \\ \boldsymbol{\delta} &= \begin{bmatrix} \boldsymbol{\Pi}_{11}' & \boldsymbol{\Pi}_{21}' \\ \boldsymbol{\Pi}_{12}' & \boldsymbol{\Pi}_{22}' \end{bmatrix} \begin{bmatrix} 1 \\ -\boldsymbol{\pi}_0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Pi}_{11}' \\ \boldsymbol{\Pi}_{12}' \end{bmatrix} \begin{bmatrix} 1 \\ -\boldsymbol{\pi}_{0,1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{bmatrix}, \\ \boldsymbol{\pi}_j &= \begin{bmatrix} \boldsymbol{\Phi}_{j,11}' & \boldsymbol{\Phi}_{j,21}' \\ \boldsymbol{\Phi}_{j,12}' & \boldsymbol{\Phi}_{j,22}' \end{bmatrix} \begin{bmatrix} 1 \\ -\boldsymbol{\pi}_0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}_{j,11}' \\ \boldsymbol{\Phi}_{j,12}' \end{bmatrix} \begin{bmatrix} 1 \\ -\boldsymbol{\pi}_{0,1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\pi}_{j,1} \\ \boldsymbol{\pi}_{j,2} \end{bmatrix}. \end{aligned} \tag{4.3.8}$$

The definitions in (4.3.8) demonstrate that, under the restriction that the errors driving $\boldsymbol{z}_{1,t}$ and $\boldsymbol{z}_{2,t}$ are uncorrelated, sparsity in the single-equation model arises when (a subset of) $\boldsymbol{z}_{2,t}$ does not Granger-Cause $\boldsymbol{z}_{1,t}$. For example, in the extreme case, where $\boldsymbol{\Pi}_{12} = \mathbf{0}$ and $\boldsymbol{\Phi}_{12} = \mathbf{0}$, it follows that $\boldsymbol{\delta}_2 = \mathbf{0}$ and $\boldsymbol{\pi}_{j,2} = 0$, respectively. Consequently, in this set-up the single-equation model reads as

$$\begin{aligned} \Delta y_t &= \boldsymbol{\delta}' \boldsymbol{z}_{t-1} + \boldsymbol{\pi}_0' \boldsymbol{x}_t + \sum_{j=1}^{p} \boldsymbol{\pi}_j' \Delta \boldsymbol{z}_{t-j} + \epsilon_{y,t} \\ &= \boldsymbol{\delta}_1' \boldsymbol{z}_{1,t-1} + \boldsymbol{\pi}_{0,1}' \Delta \boldsymbol{x}_{1,t} + \sum_{j=1}^{p} \boldsymbol{\pi}_{1,j}' \Delta \boldsymbol{z}_{1,t-j} + \epsilon_{y,t}, \end{aligned} \tag{4.3.9}$$

such that $|S_\delta| \leq N_1$ and $|S_\pi| \leq N_1(p+1)$.

Remark 4.8. The VECM in (4.3.6) can be rewritten into a non-stationary factor model with stationary idiosyncratic components, in the spirit of Banerjee et al. (2014a). Based one the VMA representation of $\boldsymbol{z}_t$ defined in (4.2.2), with $\boldsymbol{C}$ being a matrix of reduced rank, we can rewrite the process as

$$\boldsymbol{z}_t = \boldsymbol{C}\boldsymbol{s}_t + \boldsymbol{u}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{u}_t, \tag{4.3.10}$$

where $\boldsymbol{\Lambda} = \boldsymbol{B}_\perp \left(\boldsymbol{A}_\perp' \left(\boldsymbol{I} - \sum_{j=1}^{p} \boldsymbol{\Phi}_j \right) \boldsymbol{B}_\perp \right)^{-1}$, $\boldsymbol{f}_t = \boldsymbol{A}_\perp' \boldsymbol{s}_t$ and $\boldsymbol{u}_t = \boldsymbol{C}(L)\boldsymbol{\epsilon}_t + \boldsymbol{z}_0$.

Table 4.1 Dimensions, Penalties, Weights and Convergence Rates

N	p	r	$\lvert S_\delta \rvert$	$\lvert S_\pi \rvert$	k_δ	k_π	λ_R, λ_T	$\lVert \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \rVert_2$
fixed	fixed	fixed	fixed	fixed	2	1	$KT^{2/5}$	$O_p\left(T^{-1/2}\right)$
$T^{1/4}$	fixed	fixed	fixed	fixed	3	1	$KT^{2/5}$	$O_p\left(T^{-1/2}\right)$
$T^{1/4}$	$T^{1/4}$	fixed	fixed	fixed	3	2	$KT^{2/5}$	$O_p\left(T^{-1/2}\right)$
$T^{1/4}$	$T^{1/4}$	$T^{1/4}$	fixed	fixed	3	2	$KT^{2/5}$	$O_p\left(T^{-1/2}\right)$
$T^{1/4}$	$T^{1/4}$	$T^{1/4}$	fixed	$T^{1/4}$	4	2	$KT^{2/5}$	$O_p\left(T^{-3/8}\right)$
$T^{1/4}$	$T^{1/4}$	fixed	$T^{1/4}$	$T^{1/4}$	4	2	$KT^{2/5}$	$O_p\left(T^{-1/4}\right)$
$T^{1/4}$	$T^{1/4}$	$T^{1/4}$	$T^{1/4}$	$T^{1/4}$	4	2	$KT^{2/5}$	$O_p\left(T^{-3/8}\right)$

This table displays possible settings for the weights (k_δ, k_π) and penalty parameters (λ_T, λ_R) that satisfy Assumption 4.5 under a variety of asymptotic frameworks $(N, r, p, \lvert S_\delta \rvert, \lvert S_\pi \rvert)$. The convergence rate of SPECS is displayed in the last column.

This representation is particularly relevant in relation to the growth rate of $N_\delta = N - r$. Typically, the theory for consistent estimation of (4.3.10) is derived under the assumption that the N_δ factors remain fixed, while letting both N and T go to infinity. Hence, in this framework, noting that $s_\delta \leq N_\delta$, the assumptions that $\frac{s_\delta}{T^{1/4}} \to 0$ and $\frac{N_\delta}{T^{1/4}} \to 0$ in Theorems 4.1-4.3 are automatically satisfied. Consequently, the convergence rates of the initial and final estimators are given by $\lVert \hat{\boldsymbol{\gamma}}_R - \boldsymbol{\gamma} \rVert_2 = O_p\left(\sqrt{\frac{M_\pi}{T}}\right)$ and $\lVert \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \rVert_2 = O_p\left(\sqrt{\frac{s_\pi}{T}}\right)$.

The rates of convergence of $\hat{\boldsymbol{\gamma}}_R$ and $\hat{\boldsymbol{\gamma}}$, as well as the specific construction of the initial weights, are dependent on the growth rates of $N, p, r, \lvert S_\delta \rvert$ and $\lvert S_\pi \rvert$. Because of the trade-off between the dimension and the rate of convergence, the choice of the desired asymptotic framework is likely dependent on the specific application. For example, typical macro-economic applications are characterized by short panel datasets which would require a framework in which the cross-sectional dimension grows as fast as theoretically admissible. On the other hand, in applications with a large number of time series observations, such as forecasting based on high-frequency data, the assumption that the number of (potentially) relevant variables grows slow relative to the available time periods seems reasonable. Therefore, to aid interpretation of our results, we provide an overview with different asymptotic frameworks and the corresponding penalty parameters, weight constructions and convergence rates of the initial estimator in Table 4.1. The weights for δ_i and π_j are constructed as $\omega_i = \left\lvert \hat{\delta}_{R,i} \right\rvert^{-k_\delta}$ and $\omega_{N+j} = \lvert \hat{\pi}_{R,j} \rvert^{-k_\pi}$.

The first row of Table 4.1 corresponds to the classic fixed-dimensional case. It is reassuring that, similar to the OLS estimator, SPECS obtains $\sqrt{T}$-convergence, with

the additional benefit of allowing for consistent recovery of the sparsity pattern. In fact the next three rows highlight that when N, p or r diverge, while the number of relevant variables remains fixed, SPECS maintains its $\sqrt{T}$-convergence as long as the penalty weights k_δ and k_π are adjusted appropriately. In the fifth row, we allow the number of relevant stationary variables, i.e. $|S_\pi|$ to diverge as well. This setting may be preferred when the integrated time series remain persistent after being transformed to stationarity by differencing. We observe that consistency is maintained, although even sharper weights are required and the rate of convergence has reduced to $T^{-3/8}$. In the sixth row we additionally allow the number of relevant non-stationary, i.e. $|S_\delta|$, to increase, whereas the number of cointegrating vectors remains fixed. The increased number non-zero coefficients corresponding to non-stationary variables reduces the rate of convergence to $T^{-1/4}$. Interestingly, in the last row we let the dimension of the cointegrating subspace r grow at the same rate. Following Remark 4.8, this setting naturally occurs when the data is modelled by a non-stationary factor model with idiosyncratic components. In this framework, the number of stochastic trends driving the subset of relevant variables, i.e. s_δ, remains fixed, which positively affects the convergence rate of SPECS.

We consider the theoretical results presented in this section to be of a double nature. On the one hand, it is reassuring that consistent estimation remains feasible in growing dimensions and that suitable weights are available. On the other hand, we acknowledge that the required restrictions on the growth rate of the number of variables seem to caution against application of penalized regression in very high-dimensional settings. However, it is worth noting that the restrictions on N and p largely result from the use of ridge regression as an initial estimator. Indeed, the availability of a novel compatibility condition could justify the use of the lasso as an initial estimator and will allow for generalization of our theoretical results to even higher dimensional asymptotic frameworks. Accordingly, we consider this an interesting avenue for future research.

4.4 Conclusion

In this chapter, we show that SPECS may be used as an automated procedure for sparse single-equation error correction modelling in high-dimensional settings. We derive sufficient conditions under which SPECS attains simultaneous selection and estimation consistency. These results, however, strongly rely on the availability of suitable weights that aid in the identification of the subset of relevant variables. By deriving the consistency of the ridge estimator, we demonstrate how ridge regression

may be used to construct these weights, albeit under more stringent restrictions on the admissible growth rates of the irrelevant variables. On a more cautionary note, the theoretical results presented in this paper, as well as the necessary assumptions under which these results are derived, display a clear trade-off between the dimension and the estimation accuracy. This inverse relationship is more prominent in the non-stationary setting, as a result of the collinearity inducing properties of a diverging number of integrated time series.

The theoretical contributions brought forward in this chapter provide an important generalization over the fixed-dimensional case, as they justify the use of SPECS in settings in which traditional estimators perform poorly, or are rendered infeasible, as a result of the curse of dimensionality. Furthermore, we highlight several important sources through which the assumptions and asymptotic framework may be generalized even further. In particular, sharper and more direct bounds on the minimum eigenvalue of a sample covariance matrix of integrated processes can be utilized to cast SPECS into an even higher-dimensional setting. Similarly, a suitable compatibility condition can be used to validate the lasso as an initial estimator, resulting in improved weights and, again, less restrictive asymptotic frameworks. These topics remain subject to our continuing investigation.

Appendix 4.A Proofs

The proofs of our theoretical results are presented in this appendix. We start by defining several quantities of interest, some of which are simply repeated for the sake of convenience. As these quantities appear frequently throughout the remainder of the appendix, we define them here once and refer the reader to this section for a recollection of their definitions, if so needed.

First, recall that, under the assumption that $\boldsymbol{z}_0 = \boldsymbol{0}$, the moving average representation of the observed time series is given by

$$\boldsymbol{z}_t = \boldsymbol{C}\boldsymbol{s}_t + \boldsymbol{C}(L)\boldsymbol{\epsilon}_t,$$

where $\boldsymbol{C} = \boldsymbol{B}_\perp \left(\boldsymbol{A}_\perp' \left(\boldsymbol{I}_N - \sum_{j=1}^p \boldsymbol{\Phi}_j \right) \boldsymbol{B}_\perp \right)^{-1} \boldsymbol{A}_\perp'$. From this representation, one can derive the stationary processes

$$\boldsymbol{B}'\boldsymbol{z}_t = \boldsymbol{B}'\boldsymbol{C}(L)\boldsymbol{\epsilon}_t = \boldsymbol{C}^\beta(L)\boldsymbol{\epsilon}_t,$$

and

$$\Delta \boldsymbol{z}_t = \boldsymbol{C}\boldsymbol{\epsilon}_t + (1-L)\boldsymbol{C}(L)\boldsymbol{\epsilon}_t = \boldsymbol{C}^{\Delta}(L)\boldsymbol{\epsilon}_t.$$

Then, letting $\tilde{\boldsymbol{I}} = (\boldsymbol{0}, \boldsymbol{I}_{N-1})$, where $\boldsymbol{0}$ is an N-dimensional column vector of zeroes, a compact moving average representation for $\boldsymbol{w}_t = \left(\Delta \boldsymbol{x}_t', \Delta \boldsymbol{z}_{t-1}', \ldots, \Delta \boldsymbol{z}_{t-p}'\right)'$ is given by

$$\boldsymbol{w}_t = \begin{bmatrix} \tilde{\boldsymbol{I}}\boldsymbol{C}^{\Delta}(L) \\ \boldsymbol{C}^{\Delta}(L)L \\ \vdots \\ \boldsymbol{C}^{\Delta}(L)L^p \end{bmatrix} \boldsymbol{\epsilon}_t = \boldsymbol{C}^{w}(L)\boldsymbol{\epsilon}_t. \tag{4.A.1}$$

An additional useful representation follows from partitioning the data as $\boldsymbol{V} = (\boldsymbol{V}_1, \boldsymbol{V}_2)$, where $\boldsymbol{V}_1 = (\boldsymbol{Z}_{-1,S_\delta}, \boldsymbol{W}_{S_\pi})$ contains the relevant variables. In congruence with Section 4.3, the $(|S_\delta| \times r^*)$-dimensional matrix $\boldsymbol{B}_{S_\delta}$ is defined as a basis matrix for the cointegrating space of $\boldsymbol{z}_{S_\delta,t}$ and $\boldsymbol{B}_{S_\delta,\perp}$ is an $(|S_\delta| \times |S_\delta| - r^*)$-dimensional matrix for its left null space, i.e. $\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{B}_{S_\delta} = \boldsymbol{0}$. Moreover, without loss of generality, we assume that the columns of $\boldsymbol{B}_{S_\delta,\perp}$ are standardized to have unit L_1-norms. The Q-transformation is defined as

$$\begin{aligned} \boldsymbol{Q} &= \begin{bmatrix} \boldsymbol{B}_{S_\delta}' & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{|S_\pi|} \\ \boldsymbol{B}_{S_\delta,\perp}' & \boldsymbol{0} \end{bmatrix}, \\ \boldsymbol{Q}^{-1} &= \begin{bmatrix} \boldsymbol{B}_{S_\delta}\left(\boldsymbol{B}_{S_\delta}'\boldsymbol{B}_{S_\delta}\right)^{-1} & \boldsymbol{0} & \boldsymbol{B}_{S_\delta,\perp}\left(\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{B}_{S_\delta,\perp}\right)^{-1} \\ \boldsymbol{0} & \boldsymbol{I}_{|S_\pi|} & \boldsymbol{0} \end{bmatrix}, \end{aligned} \tag{4.A.2}$$

and the Q-transformed data is given by $\boldsymbol{V}_1\boldsymbol{Q}' = (\boldsymbol{Z}_{-1,S_\delta}\boldsymbol{B}_{S_\delta}, \boldsymbol{W}_{S_\pi}, \boldsymbol{Z}_{-1,S_\delta}\boldsymbol{B}_{S_\delta,\perp})$. Denote the t-th row of $\boldsymbol{V}_1\boldsymbol{Q}'$ by $\boldsymbol{v}_t = \left(\boldsymbol{v}_{1,t}', \boldsymbol{v}_{2,t}'\right)'$, where

$$\boldsymbol{v}_{1,t} = \begin{bmatrix} \boldsymbol{B}_{S_\delta}'\boldsymbol{z}_{S_\delta,t-1} \\ \boldsymbol{w}_{S_\pi,t} \end{bmatrix} = \begin{bmatrix} \boldsymbol{B}_{S_\delta}'\boldsymbol{C}(L)L \\ \boldsymbol{C}_{S_\pi}^{w}(L) \end{bmatrix} \boldsymbol{\epsilon}_t = \boldsymbol{C}^{v}(L)\boldsymbol{\epsilon}_t,$$

and

$$\boldsymbol{v}_{2,t} = \boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{z}_{S_\delta,t-1} = \boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\boldsymbol{s}_{t-1} + \boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}(L)\boldsymbol{\epsilon}_{t-1}.$$

Let $s_\pi = |S_\pi| + r^*$ and $s_\delta = |S_\delta| - r^*$ and define the scaling matrix $\boldsymbol{S}_T = \operatorname{diag}\left(\sqrt{T}\boldsymbol{I}_{s_\pi}, \frac{T}{\sqrt{s_\delta}}\boldsymbol{I}_{s_\delta}\right)$.

Then, we define the appropriately scaled sample covariance matrix as

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}_T^{-1} \left(\sum_{t=1}^{T} \boldsymbol{v}_t \boldsymbol{v}_t' \right) \boldsymbol{S}_T^{-1} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{11} & \hat{\boldsymbol{\Sigma}}_{12} \\ \hat{\boldsymbol{\Sigma}}_{21} & \hat{\boldsymbol{\Sigma}}_{22} \end{bmatrix}.$$

Based on these quantities, we proceed to describe a set of lemmas and propositions that are required for the proofs of the main theorems in this chapter.

4.A.1 Preliminary Results

In this section, we list a set of preliminary results that are used in the proofs of our main theorems in Section 4.A.2. The first result is a key ingredient for the proof of Theorem 4.1, as it explicitly describes a set on which SPECS obtains its selection consistency. The set and its sufficiency for selection consistency are derived in Proposition 1 in Zhao and Yu (2006). Accordingly, it is stated here without proof.

Proposition 4.1. *Partition* $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_{S_\gamma}', \mathbf{0}'\right)'$ *where* $\boldsymbol{\gamma}_{S_\gamma}$ *is an* s*-dimensional vector containing all non-zero coefficients and let* $\boldsymbol{v}_0 = sign(\boldsymbol{\gamma}_{S_\gamma})$*. Then,*

$$\mathbb{P}\left(\operatorname{sign}\left(\hat{\boldsymbol{\gamma}}\right) = \operatorname{sign}(\boldsymbol{\gamma})\right) \geq \mathbb{P}(\mathscr{A}_T \cap \mathscr{B}_T),$$

where

$$\mathscr{A}_T = \bigcap_{i=1}^{s} \left\{ \left| \left[\left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1} \boldsymbol{V}_1'\boldsymbol{\epsilon}_y \right]_i \right| < \left| \left[\boldsymbol{\gamma}_{S_\gamma}\right]_i \right| - \frac{1}{2}\lambda_T \left| \left[\left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1} \boldsymbol{\Omega}_1 \boldsymbol{v}_0 \right]_i \right| \right\}$$

and

$$\mathscr{B}_T = \bigcap_{i=s+1}^{N} \left\{ \left| \left[\boldsymbol{V}_2'\boldsymbol{M}\boldsymbol{\epsilon}_y\right]_i \right| < \frac{1}{2}\lambda_T \left[\left(\boldsymbol{\Omega}_2 \boldsymbol{\iota} - \left| \boldsymbol{V}_2'\boldsymbol{V}_1 \left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1} \boldsymbol{\Omega}_1 \boldsymbol{v}_0 \right| \right) \right]_i \right\},$$

with $\boldsymbol{\Omega}_1 = \operatorname{diag}(\boldsymbol{\omega}_{S_\gamma})$*,* $\boldsymbol{\Omega}_2 = \operatorname{diag}(\boldsymbol{\omega}_{S_\gamma^c})$*, and* $\boldsymbol{M} = \boldsymbol{I}_T - \boldsymbol{V}_1 \left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1} \boldsymbol{V}_1'$*.*

Next, we derive bounds on the empirical process $\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y$, which frequently appears throughout the proofs of the main results.

Lemma 4.1. *Under Assumptions 4.1-4.3, the stochastic order of the empirical process is*

$$\left\| \boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y \right\|_2 = O_p\left(s_\delta \vee \sqrt{s_\pi}\right). \tag{4.A.3}$$

Proof of Lemma 4.1. We show that $\left\| \boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y \right\|_2 = O_p\left(\left(s_\delta \vee \sqrt{s_\pi}\right)\right)$. First, note

that

$$\begin{aligned}\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 &\leq \left\|T^{-1/2}\sum_{t=1}^{T}\boldsymbol{v}_{1,t}\epsilon_{y,t}\right\|_2 + \left\|\frac{\sqrt{s_\delta}}{T}\sum_{t=1}^{T}\boldsymbol{v}_{2,t}\epsilon_{y,t}\right\|_2 \\ &\leq \left\|T^{-1/2}\sum_{t=1}^{T}\boldsymbol{C}^v(L)\boldsymbol{\epsilon}_t\epsilon_{y,t}\right\|_2 + \left\|\boldsymbol{B}'_{S_\delta,\perp}\boldsymbol{C}_{S_\delta}\left(\frac{\sqrt{s_\delta}}{T}\sum_{t=1}^{T}\boldsymbol{s}_{t-1}\epsilon_{y,t}\right)\right\|_2 \\ &\quad + \left\|\boldsymbol{B}_{S_\delta,\perp}\left(\frac{\sqrt{s_\delta}}{T}\sum_{t=1}^{T}\boldsymbol{C}_{S_\delta}(L)\boldsymbol{\epsilon}_{t-1}\epsilon_{y,t}\right)\right\|_2 =: \sum_{i=1}^{3}\|\boldsymbol{d}_i\|_2 .\end{aligned}$$

Using that

$$\mathbb{P}\left(\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}'\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 > K_\epsilon(s_\delta \vee \sqrt{s_\pi})\right) \leq \sum_{i=1}^{3}\mathbb{P}\left(\|\boldsymbol{d}_i\|_2 > \frac{K_\epsilon(s_\delta \vee \sqrt{s_\pi})}{3}\right),$$

we proceed by bounding the terms separately. First, let $\eta_{i,t} = \sum_{l=0}^{\infty}\boldsymbol{c}_{l,i}^{v\prime}\boldsymbol{\epsilon}_{t-l}$, where $\boldsymbol{c}_{l,i}^v$ is the i-th row vector of $\boldsymbol{C}_l^v$. Then, using that $\{\eta_{i,t}\epsilon_{y,t}\}$ is a martingale difference sequence, we use a combination of Markov's and Burkholder's inequality to derive

$$\begin{aligned}&\mathbb{P}\left(\|\boldsymbol{d}_1\|_2 > \frac{K_\epsilon(s_\delta \vee \sqrt{s_\pi})}{3}\right) \leq \frac{9\sum_{i=1}^{s_\pi}\mathbb{E}\left(\sum_{t=1}^{T}\eta_{i,t}\epsilon_{y,t}\right)^2}{TK_\epsilon^2(s_\delta^2 \vee s_\pi)} \\ &\leq \frac{K\sum_{i=1}^{s_\pi}\sum_{t=1}^{T}\mathbb{E}\left(\eta_{i,t}\epsilon_{y,t}\right)^2}{TK_\epsilon^2(s_\delta^2 \vee s_\pi)} \leq \frac{K^*\left(\sum_{l=0}^{\infty}\|\boldsymbol{C}_l^v\|_1\right)}{K_\epsilon^2} \leq \epsilon,\end{aligned}$$

for $K_\epsilon \geq \left(\frac{K^*\left(\sum_{l=0}^{\infty}\|\boldsymbol{C}_l^v\|_1\right)}{\epsilon}\right)^{1/2}$, where we have used that

$$\begin{aligned}\mathbb{E}\left(\eta_{i,t}\epsilon_{y,t}\right)^2 &\leq \sum_{l_1,l_2=0}^{\infty}\sum_{j_1,j_2=1}^{N}\left|c_{l_1,i,j_1}^v\right|\left|c_{l_2,i,j_2}^v\right|\mathbb{E}\left(\epsilon_{j_1,t-l_1}\epsilon_{j_2,t-l_2}\epsilon_{y,t}^2\right) \\ &\leq K\left(\sum_{l=0}^{\infty}\|\boldsymbol{C}_l^v\|_\infty\right)^2,\end{aligned}$$

by Assumption 4.1. Next, define $\boldsymbol{a}_i = \boldsymbol{C}_{S_\delta}\boldsymbol{\beta}_{S_\delta,\perp,i}$. Using the fact that $\{\boldsymbol{a}_i'\boldsymbol{s}_{t-1}\epsilon_{y,t}\}$

is a martingale difference sequence, we employ a similar strategy to show

$$\begin{aligned}
&\mathbb{P}\left(\|\boldsymbol{d}_2\|_2 > \frac{K_\epsilon(s_\delta \vee \sqrt{s_\pi})}{3}\right) \\
&\leq \frac{s_\delta 9 \sum_{i=1}^{s_\delta} \mathbb{E}\left(\sum_{t=1}^{T} \boldsymbol{a}_i' \boldsymbol{s}_{t-1}\epsilon_{y,t}\right)^2}{T^2 K_\epsilon^2 (s_\delta^2 \vee s_\pi)} \leq \frac{s_\delta K \sigma_y^2 \sum_{i=1}^{s_\delta}\sum_{t=1}^{T} \mathbb{E}\left(\boldsymbol{a}_i' \boldsymbol{s}_{t-1}\right)^2}{T^2 K_\epsilon^2 (s_\delta^2 \vee s_\pi)} \\
&\leq \frac{K \phi_{\max} \sigma_y^2 \left\|\boldsymbol{C}_{S_\delta}\right\|_2^2}{K_\epsilon^2} \leq \epsilon,
\end{aligned}$$

for $K_\epsilon \geq \left(\frac{K \phi_{\max} \sigma_y^2 \left\|\boldsymbol{C}_{S_\delta}\right\|_\infty^2}{\epsilon}\right)^{1/2}$, where we use the fact that

$$\begin{aligned}
&\mathbb{E}\left(\boldsymbol{a}_i' \boldsymbol{s}_{t-1}\right)^2 \leq \boldsymbol{a}_i' \mathbb{E}\left(\boldsymbol{s}_{t-1}\boldsymbol{s}_{t-1}'\right)\boldsymbol{a}_i = \boldsymbol{a}_i' \boldsymbol{\Sigma}_\epsilon \boldsymbol{a}_i (t-1) \\
&\leq \|\boldsymbol{a}_i\|_2^2 \, \phi_{\max}(t-1) \leq \|\boldsymbol{C}_{S_\delta}\|_2^2 \, \phi_{\max}(t-1),
\end{aligned}$$

by Assumption 4.1 and the normalization imposed on $\boldsymbol{B}_{S_\delta,\perp}$. Finally, define $\xi_{i,t} = \boldsymbol{\beta}_{S_\delta,\perp,i}' \boldsymbol{C}_{S_\delta}(L)\boldsymbol{\epsilon}_t$. Then, using that $\{\xi_{i,t}\epsilon_{y,t}\}$ is a martingale difference sequence, it follows that

$$\begin{aligned}
&\mathbb{P}\left(\|\boldsymbol{d}_3\|_2 > \frac{K_\epsilon(s_\delta \vee \sqrt{s_\pi})}{3}\right) \\
&\leq \frac{9 s_\delta \sum_{i=1}^{s_\delta} \mathbb{E}\left(\sum_{t=1}^{T} \xi_{i,t-1}\epsilon_{y,t}\right)^2}{T^2 K_\epsilon^2 (s_\delta^2 \vee s_\pi)} \leq \frac{9 s_\delta K \sigma_y^2 \sum_{i=1}^{s_\delta}\sum_{t=1}^{T} \mathbb{E}\left(\xi_{i,t-1}\right)^2}{T^2 K_\epsilon^2 (s_\delta^2 \vee s_\pi)} \\
&\leq \frac{K^* \sum_{l=0}^{\infty} \left\|\boldsymbol{C}_{S_\delta,l}\right\|_2^2}{T K_\epsilon^2} \to 0,
\end{aligned}$$

where we use that

$$\begin{aligned}
\mathbb{E}\left(\xi_{i,t-1}\right)^2 &= \sum_{l=0}^{\infty} \boldsymbol{\beta}_{S_\delta,\perp,i}' \boldsymbol{C}_{S_\delta,l} \Sigma_\epsilon \boldsymbol{C}_{S_\delta,l}' \boldsymbol{\beta}_{S_\delta,\perp,i} \\
&\leq \sum_{l=0}^{\infty} \left\|\boldsymbol{C}_{S_\delta,l}' \boldsymbol{\beta}_{S_\delta,\perp,i}\right\|_2^2 \phi_{\max} \leq \sum_{l=0}^{\infty} \left\|\boldsymbol{C}_{S_\delta,l}\right\|_2^2 \phi_{\max},
\end{aligned}$$

with $\phi_{\max}$ being the upper bound on the maximum eigenvalue of $\boldsymbol{\Sigma}_\epsilon$ from Assumption 4.1. This completes the proof. ■

Next, we proceed by deriving a minimum eigenvalue bound for the complete sample covariance matrix $\hat{\boldsymbol{\Sigma}}$. Assumption 4.4 bounds the minimum eigenvalues of the covariance matrices of the stationary and non-stationary subsets, i.e. $\hat{\boldsymbol{\Sigma}}_{11}$ and $\hat{\boldsymbol{\Sigma}}_{22}$, away from zero with probability converging to one. To translate this to a bound

on $\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}\right)$, it is necessary to complement Assumption 4.4 with a result on the off-diagonal blocks.

Lemma 4.2. *Assume that* $\frac{s_\pi}{\sqrt{T}} \to 0$ *and* $\frac{s_\delta}{T^{1/4}} \to 0$. *Then, under Assumptions 4.1-4.3, it holds that*

$$\left\|\hat{\boldsymbol{\Sigma}}_{12}\right\|_2 \xrightarrow{p} 0,$$

as $T, s_\delta, s_\pi \to \infty$.

Proof of Lemma 4.2. First, define $\boldsymbol{\eta}_t = \boldsymbol{C}^{v*}(L)\boldsymbol{\epsilon}_t$, where $\boldsymbol{C}^{v*}(L)$ is based on the Beveridge-Nelson decomposition of $\boldsymbol{C}^v(L) = \boldsymbol{C}^v(1) + \boldsymbol{C}^{v*}(L)(1-L)$. Then, with the use of summation by parts, we decompose

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_{21} &= \frac{\sqrt{s_\delta}\sum_{t=2}^T \boldsymbol{v}_{2,t}\boldsymbol{v}_{1,t}'}{T^{3/2}} = \frac{\sqrt{s_\delta}\boldsymbol{B}_{S_\delta,\perp}'\sum_{t=2}^T \boldsymbol{z}_{S_\delta,t-1}\boldsymbol{\epsilon}_t'\boldsymbol{C}^{v\prime}(L)}{T^{3/2}} \\
&= \frac{\sqrt{s_\delta}\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\sum_{t=2}^T \boldsymbol{s}_{t-1}\boldsymbol{\epsilon}_t'\boldsymbol{C}^{v\prime}(L)}{T^{3/2}} \\
&\quad + \frac{\sqrt{s_\delta}\boldsymbol{B}_{S_\delta,\perp}'\sum_{t=2}^T \boldsymbol{u}_{S_\delta,t-1}\boldsymbol{\epsilon}_t'\boldsymbol{C}^{v\prime}(L)}{T^{3/2}} \\
&= \frac{\sqrt{s_\delta}\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\sum_{t=2}^T \boldsymbol{s}_{t-1}\boldsymbol{\epsilon}_t'\boldsymbol{C}^{v\prime}(1)}{T^{3/2}} + \frac{\sqrt{s_\delta}\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\boldsymbol{s}_{T-1}\boldsymbol{\eta}_T'}{T^{3/2}} \\
&\quad + \frac{\sqrt{s_\delta}\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\sum_{t=2}^T \boldsymbol{\epsilon}_t\boldsymbol{\eta}_t'}{T^{3/2}} + \frac{\sqrt{s_\delta}\boldsymbol{B}_{S_\delta,\perp}'\sum_{t=2}^T \boldsymbol{u}_{S_\delta,t-1}\boldsymbol{\epsilon}_t'\boldsymbol{C}^{v\prime}(L)}{T^{3/2}} \\
&=: \sum_{i=1}^4 \boldsymbol{A}_i.
\end{aligned} \tag{4.A.4}$$

Hence, using that $\left\|\hat{\boldsymbol{\Sigma}}_{12}\right\|_2 \leq \sum_{i=1}^4 \|\boldsymbol{A}_i\|_2$, we proceed by showing that each $\|\boldsymbol{A}_i\|_2$ converges in probability to zero. First, let $\boldsymbol{a}_i = \boldsymbol{C}_{S_\delta}\boldsymbol{\beta}_{S_\delta,\perp,i}$ and define $\boldsymbol{b}_j = \boldsymbol{c}_j^v(1)$, where $\boldsymbol{c}_j^v(z) = \sum_{l=0}^\infty \boldsymbol{c}_{l,j}^v z^l$, with $\boldsymbol{c}_{l,j}^v$ being the j-th row of the $\boldsymbol{C}_l^v$. Note that $\{\boldsymbol{a}_i'\boldsymbol{s}_{t-1}\boldsymbol{\epsilon}_t'\boldsymbol{b}_j\}$ is a martingale difference sequence. Then, for any arbitrary constant $a > 0$, we sequentially apply Markov's, Burkholder's and the C_r-inequality to obtain

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{A}_1\|_2 \geq a\right) &\leq \frac{s_\delta\sum_{i=1}^{s_\delta}\sum_{j=1}^{s_\pi}\mathbb{E}\left(\sum_{t=2}^T \boldsymbol{a}_i'\boldsymbol{s}_{t-1}\boldsymbol{\epsilon}_t'\boldsymbol{b}_j\right)^2}{a^2T^3} \\
&\leq \frac{s_\delta K\sum_{i=1}^{s_\delta}\sum_{j=1}^{s_\pi}\sum_{t=2}^T\mathbb{E}\left(\boldsymbol{a}_i'\boldsymbol{s}_{t-1}\boldsymbol{\epsilon}_t'\boldsymbol{b}_j\right)^2}{a^2T^3} \leq \frac{s_\delta K\phi_{\max}^2\sum_{i=1}^{s_\delta}\sum_{j=1}^{s_\pi}\|\boldsymbol{a}_i\|_2^2\|\boldsymbol{b}_j\|_2^2}{a^2T} \\
&\leq \frac{s_\delta^2 s_\pi K\phi_{\max}^2\|\boldsymbol{C}_{S_\delta}\|_2^2\left(\sum_{l=0}^\infty\|\boldsymbol{C}_l^v\|_2\right)^2}{a^2T} \to 0,
\end{aligned}$$

as $\frac{s_\delta^2 s_\pi}{T} \to 0$.

Next, we focus on $\boldsymbol{A}_2$. Define $\boldsymbol{b}_{l,j} = \boldsymbol{c}_{l,j}^{v*}$ as the j-th row of $\boldsymbol{C}_l^{v*}$. Then, by sequentially applying Markov's and Minkowski's inequalities,

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{A}_2\|_2 \geq a\right) &\leq \frac{s_\delta \sum_{i=1}^{s_\delta}\sum_{j=1}^{s_\pi} \mathbb{E}\left(\sum_{l=0}^{\infty} \boldsymbol{a}_i' \boldsymbol{s}_{T-1} \boldsymbol{\epsilon}_{T-l}' \boldsymbol{b}_{l,j}\right)^2}{a^2 T^3} \\
&= \frac{s_\delta \sum_{i=1}^{s_\delta}\sum_{j=1}^{s_\pi} \mathbb{E}\left(\sum_{l=0}^{\infty}\sum_{k_1,k_2=1}^{N}\sum_{s=1}^{T-1} a_{i,k_1} b_{l,j,k_2} \epsilon_{k_1,s}\epsilon_{k_2,T-l}\right)^2}{a^2 T^3} \\
&\leq \frac{s_\delta \sum_{i=1}^{s_\delta}\sum_{j=1}^{s_\pi} \left(\sum_{l=0}^{\infty}\sum_{k_1,k_2=1}^{N}\sum_{s=1}^{T-1} |a_{i,k_1}|\,|b_{l,j,k_2}| \left(\mathbb{E}\left(\epsilon_{k_1,s}\epsilon_{k_2,T-l}\right)^2\right)^{1/2}\right)^2}{a^2 T^3} \\
&\leq \frac{s_\delta K \sum_{i=1}^{s_\delta}\sum_{j=1}^{s_\pi} \|\boldsymbol{a}_i\|_1^2 \left(\sum_{l=0}\|\boldsymbol{b}_{l,j}\|_1\right)^2}{a^2 T} \leq \frac{s_\delta^2 s_\pi K \|\boldsymbol{C}_{S_\delta}\|_\infty^2 \left(\sum_{l=0}\|\boldsymbol{C}_l^{v*}\|_\infty\right)^2}{a^2 T} \to 0.
\end{aligned}$$

Next, we focus on $\|\boldsymbol{A}_3\|_2$. Again, using a combination of Markov's and Minkowski's inequalities,

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{A}_3\|_2 \geq a\right) &\leq \frac{s_\delta \sum_{i=1}^{s_\delta}\sum_{j=1}^{s_\pi} \mathbb{E}\left(\sum_{t=1}^{T-1}\sum_{l=0}^{\infty} \boldsymbol{a}_i' \boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_{t-l}' \boldsymbol{b}_{l,j}\right)^2}{a^2 T^3} \\
&\leq \frac{s_\delta \sum_{i=1}^{s_\delta}\sum_{j=1}^{s_\pi} \left(\sum_{t=1}^{T-1}\sum_{l=0}^{\infty}\sum_{k_1,k_2=1}^{N} |a_{i,k_1}|\,|b_{l,j,k_2}| \left(\mathbb{E}\left(\epsilon_{k_1,t}\epsilon_{k_2,t-l}\right)^2\right)^{1/2}\right)^2}{a^2 T^3} \\
&\leq \frac{s_\delta K \sum_{i=1}^{s_\delta}\sum_{j=1}^{s_\pi} \|\boldsymbol{a}_i\|_1^2 \left(\sum_{l=0}\|\boldsymbol{b}_{l,j}\|_1\right)^2}{a^2 T} \leq \frac{s_\delta^2 s_\pi K \|\boldsymbol{C}_{S_\delta}\|_\infty^2 \left(\sum_{l=0}\|\boldsymbol{C}_l^{v*}\|_\infty\right)^2}{a^2 T} \to 0.
\end{aligned}$$

Finally, we consider $\|\boldsymbol{A}_4\|_2$. Define $\boldsymbol{a}_{l,i} = \boldsymbol{C}_{S_\delta,l}\boldsymbol{\beta}_{S_\delta,\perp,i}$. Then,

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{A}_4\|_2 \geq a\right) &\leq \frac{s_\delta \sum_{i=1}^{s_\delta}\sum_{j=1}^{s_\pi} \mathbb{E}\left(\sum_{t=1}^{T-1}\sum_{l_1,l_2=0}^{\infty} \boldsymbol{a}_{l_1,i}' \boldsymbol{\epsilon}_{t-1} \boldsymbol{\epsilon}_{t-l_2}' \boldsymbol{b}_{l_2,j}\right)^2}{a^2 T^3} \\
&\leq \frac{s_\delta \sum_{i=1}^{s_\delta}\sum_{j=1}^{s_\pi} \left(\sum_{t=1}^{T-1}\sum_{l_1,l_2=0}^{\infty}\sum_{k_1,k_2}^{N} |a_{l_1,i,k_1}|\,|b_{l_2,j,k_2}| \left(\mathbb{E}\left(\boldsymbol{\epsilon}_{k_1,t-1}\boldsymbol{\epsilon}_{k_2,t-l_2}\right)^2\right)^{1/2}\right)^2}{a^2 T^3} \\
&\leq \frac{s_\delta K \sum_{i=1}^{s_\delta}\sum_{j=1}^{s_\pi} \left(\sum_{l=0}^{\infty}\|\boldsymbol{a}_{l,i}\|_1\right)^2 \left(\sum_{l=0}^{\infty}\|\boldsymbol{b}_{l,j}\|_1\right)^2}{a^2 T} \\
&\leq \frac{s_\delta^2 s_\pi K \left(\sum_{l=0}^{\infty}\|\boldsymbol{C}_{S_\delta,l}\|_\infty\right)^2 \left(\sum_{l=0}^{\infty}\|\boldsymbol{C}_l^{v*}\|_\infty\right)^2}{a^2 T} \to 0,
\end{aligned}$$

thereby completing the argument. ■

Combining Assumption 4.4 with Lemma 4.2, we obtain the following immediate

result.

Corollary 4.1. *Under the same assumption as in Lemma 4.2, there exists a constant* $\phi^* > 0$*, such that*

$$\mathbb{P}\left(\lambda_1\left(\hat{\boldsymbol{\Sigma}}\right) \geq \phi^*\right) \to 1,$$

as $T, s_\delta, s_\pi \to \infty$.

Proof of Corollary 4.1. Let $\tilde{\boldsymbol{\Sigma}} = \operatorname{diag}\left(\hat{\boldsymbol{\Sigma}}_{11}, \hat{\boldsymbol{\Sigma}}_{22}\right)$ and $\check{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}} - \tilde{\boldsymbol{\Sigma}}$. Note that

$$\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}\right) \geq \lambda_{\min}\left(\tilde{\boldsymbol{\Sigma}}\right) + \lambda_{\min}\left(\check{\boldsymbol{\Sigma}}\right) \geq \lambda_{\min}\left(\tilde{\boldsymbol{\Sigma}}\right) - \left\|\check{\boldsymbol{\Sigma}}\right\|_2.$$

Furthermore,

$$\begin{aligned}\mathbb{P}\left(\lambda_{\min}\left(\tilde{\boldsymbol{\Sigma}}\right) < \phi\right) &= \mathbb{P}\left(\min\left(\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{11}\right), \lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{22}\right)\right) < \phi\right) \\ &\leq \mathbb{P}\left(\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{11}\right) < \phi\right) + \mathbb{P}\left(\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{22}\right) < \phi\right) \to 0.\end{aligned}$$

Thus, for any $\epsilon > 0$, we may choose a T_1 such that $\mathbb{P}\left(\lambda_{\min}\left(\tilde{\boldsymbol{\Sigma}}\right) < \phi\right) \leq \frac{\epsilon}{2}$ for all $T > T_1$. Moreover, by Lemma 4.2, there exists a T_2 such that $\mathbb{P}\left(\left\|\check{\boldsymbol{\Sigma}}\right\|_2 \geq \frac{\phi}{2}\right) \leq \frac{\epsilon}{2}$, whenever $T > T_2$. Then, for all $T > \max(T_1, T_2)$, we have

$$\begin{aligned}&\mathbb{P}\left(\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}\right) < \frac{\phi}{2}\right) \leq \mathbb{P}\left(\lambda_{\min}\left(\tilde{\boldsymbol{\Sigma}}\right) - \left\|\check{\boldsymbol{\Sigma}}\right\|_2 < \frac{\phi}{2}\right) \\ &\leq \mathbb{P}\left(\lambda_{\min}\left(\tilde{\boldsymbol{\Sigma}}\right) - \left\|\check{\boldsymbol{\Sigma}}\right\|_2 < \frac{\phi}{2}, \left\|\check{\boldsymbol{\Sigma}}\right\|_2 < \frac{\phi}{2}\right) + \mathbb{P}\left(\left\|\check{\boldsymbol{\Sigma}}\right\|_2 \geq \frac{\phi}{2}\right) \\ &\leq \mathbb{P}\left(\lambda_{\min}\left(\tilde{\boldsymbol{\Sigma}}\right) < \phi\right) + \frac{\epsilon}{2} \leq \epsilon.\end{aligned}$$

Since ϵ was chosen arbitrarily, the claim is shown for $\phi^* = \frac{\phi}{2}$. The same proof works for any $0 < \phi^* < \phi$. ■

Finally, we note that Lemma 4.1 and Corollary 4.1 have natural counterparts based on the full dataset. This is described in the following corollary.

Corollary 4.2. *Let* $N_\delta = N - r$ *and* $M_\pi = M + r$ *and define the scaling and rotation matrices* $\boldsymbol{S}_R = \operatorname{diag}\left(\sqrt{T}\boldsymbol{I}_{M_\pi}, \frac{T}{N_\delta}\boldsymbol{I}_{N_\delta}\right)$ *and*

$$\boldsymbol{Q}_R = \begin{bmatrix} \left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1/2}\boldsymbol{B}' & 0 \\ 0 & \boldsymbol{I}_M \\ \left(\boldsymbol{B}_\perp'\boldsymbol{B}_\perp\right)^{-1/2}\boldsymbol{B}_\perp' & 0 \end{bmatrix}.$$

Table 4.2 List of conversions

Old	s_δ	s_π	$\boldsymbol{B}_{S_\delta}$	$\boldsymbol{B}_{S_\delta,\perp}$	$\boldsymbol{Q}$	$\boldsymbol{v}_t$	$\boldsymbol{v}_{1,t}$	$\boldsymbol{v}_{2,t}$	$\boldsymbol{S}_T$
New	N_δ	M_π	$\boldsymbol{B}$	$\boldsymbol{B}_\perp$	$\boldsymbol{Q}_R$	$\boldsymbol{v}_{R,t}$	$\boldsymbol{v}_{R1,t}$	$\boldsymbol{v}_{R2,t}$	$\boldsymbol{S}_R$

This table lists the conversions necessary to apply the proofs of Lemmas 4.1-4.2 and Corollary 4.1.

Furthermore, define $\hat{\boldsymbol{\Sigma}}_R = \boldsymbol{S}_R^{-1}\boldsymbol{Q}_R\boldsymbol{V}'\boldsymbol{Q}_R'\boldsymbol{S}_R^{-1}$. *Assume that* $\frac{N_\delta}{T^{1/4}} \to 0$ *and* $\frac{M_\pi}{\sqrt{T}} \to 0$. *Then, under Assumptions 4.1-4.3 and 4.6,*

1. $\mathbb{P}\left(\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_R\right) \geq \phi_R\right) \to 1$, *as* $T, N_\delta, M_\pi \to \infty$, *and*
2. $\left\|\boldsymbol{S}_R^{-1}\boldsymbol{Q}_R\boldsymbol{V}'\boldsymbol{\epsilon}_y\right\|_2 = O_p\left(N_\delta \vee \sqrt{M_\pi}\right)$.

Proof. First, note that $\boldsymbol{Q}_R'\boldsymbol{Q}_R = \boldsymbol{I}_{M+N}$ by construction. From the VMA representation (4.2.2), it follows that $\boldsymbol{V}\boldsymbol{Q}_R' = \left[\boldsymbol{V}_{R1}, \boldsymbol{V}_{R2}\right]$, where $\boldsymbol{V}_{R1}$ is an $(T \times M_\pi)$-dimensional matrix containing stationary processes and $\boldsymbol{V}_{R2}$ is an $(T \times N_\delta)$-dimensional matrix containing integrated processes. We denote the rows of $\boldsymbol{V}_{R1}$ and $\boldsymbol{V}_{R2}$ by $\boldsymbol{v}_{R1,t}$ and $\boldsymbol{v}_{R2,t}$, respectively. Then, after a set of suitable replacements, the proof of Corollary 4.2 is entirely analogues to the proofs of Lemma 4.1-4.2 and Corollary 4.1. The required substitutions are summarized in Table 4.2.

■

4.A.2 Main Theorems

Proof of Theorem 4.1. Based on Proposition 4.1, it is suffices to show that $\mathbb{P}(\mathscr{A}_T \cap \mathscr{B}_T) \to 1$ as $T, N \to \infty$ or, equivalently, that $\mathbb{P}(\mathscr{A}_T^c) \to 0$ and $\mathbb{P}(\mathscr{B}_T^c) \to 0$. Thus, we start by deriving that $\mathbb{P}(\mathscr{A}_T^c) \to 0$.

Recall the definitions of $\boldsymbol{S}_T = \text{diag}\left(\sqrt{T}\boldsymbol{I}_{s_\pi}, \frac{T}{\sqrt{s_\delta}}\boldsymbol{I}_{s_\delta}\right)$ and define $\boldsymbol{Q}$ as in (4.A.2), with $\|\boldsymbol{Q}\|_\infty \leq 1$ by the normalization on $\boldsymbol{B}_{S_\delta}$ and $\boldsymbol{B}_{S_\delta,\perp}$. Then, for T large enough,

we may write the set $\mathscr{A}_T^c$ as

$$\begin{aligned}
\mathscr{A}_T^c &= \bigcup_{i=1}^{s} \Big\{ \Big| \Big[\boldsymbol{Q}' \boldsymbol{S}_T^{-1} \left(\boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{V}_1 \boldsymbol{Q}' \boldsymbol{S}_T^{-1} \right)^{-1} \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \Big]_i \Big| \\
&\quad \geq |\gamma_{S_\gamma,i}| - \frac{1}{2}\lambda_T \Big| \Big[\boldsymbol{Q}' \boldsymbol{S}_T^{-1} \left(\boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{V}_1 \boldsymbol{Q}' \boldsymbol{S}_T^{-1} \right)^{-1} \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{\Omega}_1 \boldsymbol{v}_0 \Big]_i \Big| \Big\} \\
&= \bigcup_{i=1}^{s} \Big\{ \Big| \Big[\boldsymbol{Q}' \boldsymbol{S}_T^{-1} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \Big]_i \Big| \\
&\quad \geq |\gamma_{S_\gamma,i}| - \frac{1}{2}\lambda_T \Big| \Big[\boldsymbol{Q}' \boldsymbol{S}_T^{-1} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{\Omega}_1 \boldsymbol{v}_0 \Big]_i \Big| \Big\} \\
&\subseteq \Big\{ \Big\| \boldsymbol{Q}' \boldsymbol{S}_T^{-1} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \Big\|_2 \\
&\quad \geq \min_{1 \leq i \leq s} |\gamma_{S_\gamma,i}| - \frac{1}{2}\lambda_T \Big\| \boldsymbol{Q}' \boldsymbol{S}_T^{-1} \boldsymbol{Q}' \boldsymbol{S}_T^{-1} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{\Omega}_1 \boldsymbol{v}_0 \Big\|_2 \Big\}
\end{aligned} \tag{4.A.5}$$

We proceed by bounding the three quantities in (4.A.5) separately. First, by our assumption on the growth rate of s_δ in Theorem 4.1,

$$\frac{s_\delta}{T} \leq \frac{1}{\sqrt{T}} \Rightarrow \left\| \boldsymbol{S}_T^{-1} \right\|_2 = \frac{1}{\sqrt{T}},$$

for large enough T. Moreover, letting $s = (s_\delta \vee s_\pi)$,

$$\left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{\Omega}_1 \boldsymbol{v}_0 \right\|_2 \leq \left\| \boldsymbol{S}_T^{-1} \right\|_2 \|\boldsymbol{Q}\|_2 \|\boldsymbol{\Omega}_1\|_2 \|\boldsymbol{v}_0\|_2 \leq \frac{2\sqrt{s}}{T^{1/2-\xi}}.$$

Then, on a set with probability converging to one,

$$\begin{aligned}
\left\| \boldsymbol{Q}' \boldsymbol{S}_T^{-1} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \right\|_2 &\leq \left\| \boldsymbol{S}_T^{-1} \right\|_2 \|\boldsymbol{Q}\|_2 \left\| \hat{\boldsymbol{\Sigma}}^{-1} \right\|_2 \left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \right\|_2 \\
&\leq \frac{\left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \right\|_2}{\sqrt{T}\phi}.
\end{aligned} \tag{4.A.6}$$

Furthermore, on the same set,

$$\begin{aligned}
&\left\| \boldsymbol{Q}' \boldsymbol{S}_T^{-1} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{\Omega}_1 \boldsymbol{v}_0 \right\|_2 \\
&\quad \leq \left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \right\|_2 \left\| \hat{\boldsymbol{\Sigma}}^{-1} \right\|_2 \left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{\Omega}_1 \boldsymbol{v}_0 \right\|_2 \leq \frac{2\sqrt{s}}{\phi T^{1-\xi}}.
\end{aligned} \tag{4.A.7}$$

Based on (4.A.6) and (4.A.7), we obtain probability bounds for $\mathscr{A}_T^c$ as follows:

$$\begin{aligned}\mathbb{P}(\mathscr{A}_T^c) &\leq \mathbb{P}\left(\left\|\boldsymbol{Q}'\boldsymbol{S}_T^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2\right. \\ &\qquad \left.\geq |\gamma_{\min}| - \lambda_T\left\|\boldsymbol{Q}'\boldsymbol{S}_T^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{\Omega}_1\boldsymbol{v}_0\right\|_2\right) \\ &\leq \mathbb{P}\left(\frac{\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2}{\sqrt{T}\phi} \geq |\gamma_{\min}| - \frac{2\lambda_T\sqrt{s}}{\phi T^{1-\xi}}\right) + o(1) \\ &= \mathbb{P}\left(\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 \geq \phi\,|\gamma_{\min}|\,\sqrt{T} - \frac{2\lambda_T\sqrt{s}}{T^{1/2-\xi}}\right) + o(1).\end{aligned} \tag{4.A.8}$$

Then, to establish that $\mathbb{P}(\mathscr{A}_T^c) \to 0$, by Lemma 4.1 it suffices that

$$\frac{|\gamma_{\min}|\,\sqrt{T}}{(s_\delta \vee \sqrt{s_\pi})} \to \infty, \text{ and } \frac{|\gamma_{\min}|\,T^{1-\xi}}{\lambda_T\sqrt{s}} \to \infty. \tag{4.A.9}$$

The first condition in (4.A.9) correspond to part 1 of Assumption 4.5. Regarding the second condition, for sufficiently large T,

$$\frac{|\gamma_{\min}|\,T^{1-\xi}}{\lambda_T\sqrt{s}} \geq \frac{\left(s_\delta \vee \sqrt{s_\pi}\right)T^{1/2-\xi}}{\lambda_T\sqrt{s}} \to \infty,$$

where the divergence follows from part 2 of Assumption 4.5. Hence, we conclude that $\mathbb{P}(\mathscr{A}_T^c) \to 0$.

Next, we show that $\mathbb{P}(\mathscr{B}_T^c) \to 0$. Recall from Proposition 4.1,

$$\begin{aligned}\mathscr{B}_T^c &= \bigcup_{i=s+1}^{N}\left\{\left|[\boldsymbol{V}_2'\boldsymbol{M}\boldsymbol{\epsilon}_y]_i\right| \geq \frac{\lambda_T}{2}\left[\left(\boldsymbol{\Omega}_2\boldsymbol{\iota} - \left|\boldsymbol{V}_2'\boldsymbol{V}_1\left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1}\boldsymbol{\Omega}_1\boldsymbol{v}_0\right|\right)\right]_i\right\} \\ &= \left(\bigcup_{i=1}^{|S_\delta^c|}\left\{\left|\boldsymbol{z}_{S_\delta^c,i}'\boldsymbol{M}\boldsymbol{\epsilon}_y\right| \geq \frac{\lambda_T}{2}\omega_{S_\delta^c,i} - \frac{\lambda_T}{2}\left|\boldsymbol{z}_{S_\delta^c,i}'\boldsymbol{V}_1\left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1}\boldsymbol{\Omega}_1\boldsymbol{v}_0\right|\right\}\right) \bigcup \\ &\quad \left(\bigcup_{i=1}^{|S_\pi^c|}\left\{\left|\boldsymbol{w}_{S_\pi^c,i}'\boldsymbol{M}\boldsymbol{\epsilon}_y\right| \geq \frac{\lambda_T}{2}\omega_{S_\pi^c,i} - \frac{\lambda_T}{2}\left|\boldsymbol{w}_{S_\pi^c,i}'\boldsymbol{V}_1\left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1}\boldsymbol{\Omega}_1\boldsymbol{v}_0\right|\right\}\right) \\ &= \mathscr{B}_{z,T}^c \cup \mathscr{B}_{w,T}^c.\end{aligned} \tag{4.A.10}$$

Focussing first on $\mathscr{B}_{z,T}^c$,

$$\mathscr{B}_{z,T}^c \subseteq \left\{\left\|\boldsymbol{Z}_{S_\delta^c}'\boldsymbol{M}\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\lambda_T}{2}\omega_{S_\delta^c,\min} - \frac{\lambda_T}{2}\left\|\boldsymbol{Z}_{S_\delta^c}'\boldsymbol{V}_1\left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1}\boldsymbol{\Omega}_1\boldsymbol{v}_0\right\|_2\right\} \tag{4.A.11}$$

We proceed by bounding each individual term in (4.A.11). First, on a set with

probability converging to one,

$$\begin{aligned}\left\|\boldsymbol{Z}'_{S_\delta^c}\boldsymbol{M}\boldsymbol{\epsilon}_y\right\|_2 &\leq \left\|\boldsymbol{Z}'_{S_\delta^c}\boldsymbol{\epsilon}_y\right\|_2 + \left\|\boldsymbol{Z}'_{S_\delta^c}\boldsymbol{V}_1\left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 \\ &\leq \left\|\boldsymbol{Z}'_{S_\delta^c}\boldsymbol{\epsilon}_y\right\|_2 + \frac{\left\|\boldsymbol{Z}_{S_\delta^c}\right\|_2}{\sqrt{\phi}}\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2,\end{aligned} \tag{4.A.12}$$

where the last inequality follows from the fact that

$$\begin{aligned}\left\|\boldsymbol{V}_1\left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 &= \left(\boldsymbol{\epsilon}_y'\boldsymbol{V}_1\left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right)^{1/2} \\ &= \left(\boldsymbol{\epsilon}_y'\boldsymbol{V}_1\boldsymbol{Q}'\boldsymbol{S}_T^{-1}\left(\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{V}_1\boldsymbol{Q}'\boldsymbol{S}_T^{-1}\right)^{-1}\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right)^{1/2} \\ &= \left\|\left(\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{V}_1\boldsymbol{Q}'\boldsymbol{S}_T^{-1}\right)^{-1/2}\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 \leq \frac{\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2}{\sqrt{\phi}}\end{aligned}$$

by Corollary 4.1. By the same argument, it follows that

$$\left\|\boldsymbol{Z}'_{S_\delta^c}\boldsymbol{V}_1\left(\boldsymbol{V}_1'\boldsymbol{V}_1\right)^{-1}\boldsymbol{\Omega}_1\boldsymbol{v}_0\right\|_2 \leq \frac{\left\|\boldsymbol{Z}_{S_\delta^c}\right\|_2}{\sqrt{\phi}}\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{\Omega}_1\boldsymbol{v}_0\right\|_2 \leq \frac{2\sqrt{s}\left\|\boldsymbol{Z}_{S_\delta^c}\right\|_2}{\sqrt{\phi}T^{1/2-\xi}}. \tag{4.A.13}$$

Then, plugging (4.A.12) and (4.A.13) into (4.A.11), we obtain

$$\begin{aligned}\mathbb{P}\left(\mathscr{B}_{z,T}^c\right) \leq\ &\mathbb{P}\left(\left\|\boldsymbol{Z}'_{S_\delta^c}\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\lambda_T\omega_{S_\delta^c,\min}}{4} - \frac{\lambda_T\sqrt{s}\left\|\boldsymbol{Z}_{S_\delta^c}\right\|_2}{2\sqrt{\phi}T^{1/2-\xi}}\right) \\ &+ \mathbb{P}\left(\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\sqrt{\phi}\lambda_T\omega_{S_\delta^c,\min}}{4\left\|\boldsymbol{Z}_{S_\delta^c}\right\|_2} - \frac{\lambda_T\sqrt{s}}{2T^{1/2-\xi}}\right).\end{aligned} \tag{4.A.14}$$

We proceed by deriving the stochastic order of the common term $\left\|\boldsymbol{Z}_{S_\delta^c}\right\|_2$, by noting that,

$$\begin{aligned}\mathbb{P}\left(\left\|T^{-1}N^{-1/2}\boldsymbol{Z}_{S_\delta^c}\right\|_2 \geq a\right) \leq\ &\mathbb{P}\left(\left\|\boldsymbol{C}_{S_\delta^c}\right\|_2\left\|T^{-1}N^{-1/2}\boldsymbol{S}\right\|_2 \geq \frac{K_\epsilon}{2}\right) \\ &+ \mathbb{P}\left(\left\|T^{-1}N^{-1/2}\boldsymbol{U}_{S_\delta^c}\right\|_2 \geq \frac{K_\epsilon}{2}\right).\end{aligned}$$

Furthermore, by Markov's inequality and Assumption 4.1,

$$\begin{aligned}\mathbb{P}\left(\left\|\boldsymbol{C}_{S_\delta^c}\right\|_2\left\|T^{-1}N^{-1/2}\boldsymbol{S}\right\|_2 \geq \frac{K_\epsilon}{2}\right) &\leq \frac{4\left\|\boldsymbol{C}_{S_\delta^c}\right\|_2^2\sum_{i=1}^N\sum_{t=1}^{T-1}\mathbb{E}\left(s_{i,t}\right)^2}{K_\epsilon^2T^2N} \\ &\leq \frac{4\left\|\boldsymbol{C}_{S_\delta^c}\right\|_2^2K}{K_\epsilon^2} \leq \epsilon,\end{aligned}$$

for $K_\epsilon \geq \left(\frac{4 \left\| \boldsymbol{C}_{S_\delta^c} \right\|_2^2 K}{\epsilon} \right)^{1/2}$, and

$$\begin{aligned}
\mathbb{P}\left(\left\| T^{-1} N^{-1/2} \boldsymbol{U}_{S_\delta^c} \right\|_2 \geq \frac{K_\epsilon}{2} \right) &\leq \frac{4 \sum_{i=1}^{|S_\delta^c|} \sum_{t=1}^{T-1} \mathbb{E}\left(u_{S_\delta^c,i,t} \right)^2}{K_\epsilon^2 T^2 N} \\
&\leq \frac{4\phi_{\max} \sum_{i=1}^{|S_\delta^c|} \sum_{l=0}^{\infty} \left\| \boldsymbol{c}_{S_\delta^c,l,i} \right\|_2^2}{K_\epsilon^2 T N} \leq \frac{4\phi_{\max} \sum_{l=0}^{\infty} \left\| \boldsymbol{C}_{S_\delta^c,l} \right\|_2^2}{K_\epsilon^2 T} \leq \frac{K}{T} \to 0.
\end{aligned}$$

Hence, $\left\| \boldsymbol{Z}_{S_\delta^c} \right\|_2 = O_p\left(T\sqrt{N} \right)$, i.e. for all $\epsilon > 0$ there exist $K_\epsilon, T^*, N^* > 0$ such that $\mathbb{P}\left(\left\| \boldsymbol{Z}_{S_\delta^c} \right\|_2 \geq T\sqrt{N} K_\epsilon \right) \leq \epsilon$ for all $T > T^*$ and $N > N^*$. We use this to simplify the two RHS terms of (4.A.14).

For sufficiently large T, the first RHS term of (4.A.14) is bounded by

$$\begin{aligned}
&\mathbb{P}\left(\left\| \boldsymbol{Z}'_{S_\delta^c} \boldsymbol{\epsilon}_y \right\|_2 \geq \frac{\lambda_T \omega_{S_\delta^c,\min}}{4} - \frac{\lambda_T \sqrt{s} \left\| \boldsymbol{Z}_{S_\delta^c} \right\|_2}{2\sqrt{\phi} T^{1/2-\xi}} \right) \\
&\leq \mathbb{P}\left(\left\| \boldsymbol{C}_{S_\delta^c} \boldsymbol{S}' \boldsymbol{\epsilon}_y \right\|_2 \geq \frac{\lambda_T \omega_{S_\delta^c,\min}}{8} - \frac{\lambda_T K_\epsilon \sqrt{s} T^{1/2+\xi} \sqrt{N}}{4\sqrt{\phi}} \right) \\
&\quad + \mathbb{P}\left(\left\| \boldsymbol{U}'_{S_\delta^c} \boldsymbol{\epsilon}_y \right\|_2 \geq \frac{\lambda_T \omega_{S_\delta^c,\min}}{8} - \frac{\lambda_T K_\epsilon \sqrt{s} T^{1/2+\xi} \sqrt{N}}{4\sqrt{\phi}} \right) + \epsilon.
\end{aligned} \tag{4.A.15}$$

Then, using that $\{ s_{i,t-1} \epsilon_{y,t} \}$ is a m.d.s., it follows from application of Burkholder's inequality in combination with the C_r-inequality, that for an $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}\left(\frac{\left\| \boldsymbol{C}_{S_\delta^c} \boldsymbol{S}' \boldsymbol{\epsilon}_y \right\|_2}{T\sqrt{N}} \geq K_\epsilon \right) &\leq \frac{\left\| \boldsymbol{C}_{S_\delta^c} \right\|_2^2 \sum_{i=1}^{N} \mathbb{E}\left(\sum_{t=2}^{T} s_{i,t-1} \epsilon_{y,t} \right)^2}{K_\epsilon^2 T^2 N} \\
&\leq \frac{K \left\| \boldsymbol{C}_{S_\delta^c} \right\|_2^2 \sigma_y^2 \sum_{i=1}^{N} \sum_{t=1}^{T-1} \mathbb{E}(s_{i,t-1})^2}{K_\epsilon^2 T^2 N} \leq \frac{K^* \left\| \boldsymbol{C}_{S_\delta^c} \right\|_2^2 \sigma_y^2}{K_\epsilon^2} \leq \epsilon,
\end{aligned} \tag{4.A.16}$$

for $K_\epsilon \geq \left(\frac{K^* \left\| \boldsymbol{C}_{S_\delta^c} \right\|_2^2 \sigma_y^2}{\epsilon} \right)^{1/2}$, and

$$\begin{aligned}
&\mathbb{P}\left(\frac{\left\| \boldsymbol{U}'_{S_\delta^c} \boldsymbol{\epsilon}_y \right\|_2}{T\sqrt{N}} \geq K_\epsilon \right) \leq \frac{\sum_{i=1}^{|S_\delta^c|} \mathbb{E}\left(\sum_{t=2}^{T} \sum_{l=0}^{\infty} \boldsymbol{c}'_{S_\delta,l,i} \boldsymbol{\epsilon}_{t-1-l} \epsilon_{y,t} \right)^2}{K_\epsilon^2 T^2 N} \\
&\leq \frac{K\sigma_y^2 \sum_{i=1}^{|S_\delta^c|} \sum_{t=2}^{T} \sum_{l=0}^{\infty} \mathbb{E}\left(\boldsymbol{c}'_{S_\delta,l,i} \boldsymbol{\epsilon}_{t-1-l} \right)^2}{K_\epsilon^2 T^2 N} \\
&\leq \frac{K\sigma_y^2 \phi_{\max} \sum_{i=1}^{|S_\delta^c|} \sum_{l=0}^{\infty} \left\| \boldsymbol{c}_{S_\delta,l,i} \right\|_2^2}{K_\epsilon^2 T N} \leq \frac{K\sigma_y^2 \phi_{\max} \sum_{l=0}^{\infty} \left\| \boldsymbol{C}_{S_\delta,l} \right\|_2^2}{K_\epsilon^2 T N} \to 0.
\end{aligned} \tag{4.A.17}$$

Hence, based on (4.A.16) and (4.A.17),

$$\mathbb{P}\left(\left\| \boldsymbol{Z}'_{S_\delta^c} \boldsymbol{\epsilon}_y \right\|_2 \geq \frac{\lambda_T \omega_{S_\delta^c,\min}}{4} - \frac{\lambda_T \sqrt{s} \left\| \boldsymbol{Z}_{S_\delta^c} \right\|_2}{2\sqrt{\phi} T^{1/2-\xi}} \right) \to 0.$$

if

$$\frac{\omega_{S_\delta^c,\min}}{\sqrt{s} T^{1/2+\xi} \sqrt{N}} \to \infty, \text{ and } \frac{\lambda_T \omega_{S_\delta^c,\min}}{T\sqrt{N}} \to \infty. \tag{4.A.18}$$

Both conditions in (4.A.18) are satisfied under Assumption 4.5.

Next, we focus on the second RHS term of (4.A.14). First, again using that $\left\| \boldsymbol{Z}_{S_\delta^c} \right\|_2 = O_p(T\sqrt{N})$, it holds that

$$\begin{aligned}
&\mathbb{P}\left(\left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \right\|_2 \geq \frac{\sqrt{\phi} \lambda_T \omega_{S_\delta^c,\min}}{4 \left\| \boldsymbol{Z}_{S_\delta^c} \right\|_2} - \frac{\lambda_T \sqrt{s}}{2T^{1/2-\xi}} \right) \\
&\leq \mathbb{P}\left(\left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \right\|_2 \geq \frac{\sqrt{\phi} \lambda_T \omega_{S_\delta^c,\min}}{4K_\epsilon T\sqrt{N}} - \frac{\lambda_T \sqrt{s}}{2T^{1/2-\xi}} \right) + \epsilon.
\end{aligned} \tag{4.A.19}$$

Then, based on Lemma 4.1, for the RHS of (4.A.19) to converge to zero, it is sufficient that

$$\frac{\omega^c_{S_\delta,\min}}{\sqrt{s} T^{1/2+\xi} \sqrt{N}} \to \infty, \text{ and } \frac{\lambda_T \omega_{S_\delta^c,\min}}{(s_\delta \vee \sqrt{s_\pi}) T\sqrt{N}} \to \infty.$$

Both conditions are satisfied under Assumption 4.5. Consequently, both RHS terms of (4.A.14) converge to zero, thereby concluding that $\mathbb{P}\left(\mathscr{B}^c_{z,T}\right) \to 0$.

The last remaining part in proving Theorem 4.1 is to show that $\mathbb{P}\left(\mathscr{B}^c_{w,T}\right) \to 0$,

where $\mathscr{B}^c_{w,T}$ is defined in (4.A.10). First, note that

$$\mathscr{B}^c_{w,T} \subseteq \left\{ \left\| \boldsymbol{W}'_{S^c_\pi} \boldsymbol{M} \boldsymbol{\epsilon}_y \right\|_2 \geq \frac{\lambda_T}{2} \omega_{S^c_\pi,\min} - \frac{\lambda_T}{2} \left\| \boldsymbol{W}'_{S^c_\pi} \boldsymbol{V}_1 \left(\boldsymbol{V}_1' \boldsymbol{V}_1 \right)^{-1} \boldsymbol{\Omega}_1 \boldsymbol{v}_0 \right\| \right\}.$$

Furthermore, on a set with probability converging to one,

$$\left\| \boldsymbol{W}'_{S^c_\pi} \boldsymbol{M} \boldsymbol{\epsilon}_y \right\|_2 \leq \left\| \boldsymbol{W}'_{S^c_\pi} \boldsymbol{\epsilon}_y \right\|_2 + \frac{\left\| \boldsymbol{W}_{S^c_\pi} \right\|_2 \left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \right\|_2}{\sqrt{\phi}} \tag{4.A.20}$$

and

$$\left\| \boldsymbol{W}'_{S^c_\pi} \boldsymbol{V}_1 \left(\boldsymbol{V}_1' \boldsymbol{V}_1 \right)^{-1} \boldsymbol{\Omega}_1 \boldsymbol{v}_0 \right\|_2 \leq \frac{\left\| \boldsymbol{W}_{S^c_\pi} \right\|_2 \left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{\Omega}_1 \boldsymbol{v}_0 \right\|}{\sqrt{\phi}} \leq \frac{2\sqrt{s} \left\| \boldsymbol{W}_{S^c_\pi} \right\|_2}{\sqrt{\phi} T^{1/2-\xi}}. \tag{4.A.21}$$

Then, plugging (4.A.20)-(4.A.21) into $\mathscr{B}^c_{w,T}$ from (4.A.10), we obtain

$$\begin{aligned}
\mathbb{P}\left(\mathscr{B}^c_{w,T}\right) &\leq \mathbb{P}\left(\left\| \boldsymbol{W}'_{S^c_\pi} \boldsymbol{\epsilon}_y \right\|_2 + \frac{\left\| \boldsymbol{W}_{S^c_\pi} \right\|_2 \left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \right\|_2}{\sqrt{\phi}} \right. \\
&\qquad \left. \geq \frac{\lambda_T \omega_{S^c_\pi,\min}}{2} - \frac{\lambda_T \sqrt{s} \left\| \boldsymbol{W}_{S^c_\pi} \right\|_2}{\sqrt{\phi} T^{1/2-\xi}} \right) + o(1) \\
&\leq \mathbb{P}\left(\left\| \boldsymbol{W}'_{S^c_\pi} \boldsymbol{\epsilon}_y \right\|_2 \geq \frac{\lambda_T \omega_{S^c_\pi,\min}}{4} - \frac{\lambda_T \sqrt{s} \left\| \boldsymbol{W}_{S^c_\pi} \right\|_2}{2\sqrt{\phi} T^{1/2-\xi}} \right) \\
&\quad + \mathbb{P}\left(\left\| \boldsymbol{S}_T^{-1} \boldsymbol{Q} \boldsymbol{V}_1' \boldsymbol{\epsilon}_y \right\|_2 \geq \frac{\lambda_T \sqrt{\phi} \omega_{S^c_\pi,\min}}{4 \left\| \boldsymbol{W}_{S^c_\pi} \right\|_2} - \frac{\lambda_T \sqrt{s}}{2T^{1/2-\xi}} \right) + o(1) \\
&= \mathbb{P}\left(\mathscr{B}^c_{w_1,T}\right) + \mathbb{P}\left(\mathscr{B}^c_{w_2,T}\right) + o(1).
\end{aligned} \tag{4.A.22}$$

Next, we derive the stochastic order of the common term $\left\| \boldsymbol{W}_{S^c_\pi} \right\|_2$. Recalling that $\boldsymbol{w}_{i,t} = \sum_{l=0}^{\infty} \boldsymbol{c}_{l,i}^{w\prime} \boldsymbol{\epsilon}_{t-l}$, it holds that

$$\mathbb{E}\left(w_{i,t}\right)^2 = \sum_{l=0}^{\infty} \boldsymbol{c}_{l,i}^{w\prime} \boldsymbol{\Sigma}_\epsilon \boldsymbol{c}_{l,i}^{w} \leq \phi_{\max} \sum_{l=0}^{\infty} \left\| \boldsymbol{c}_{l,i}^{w} \right\|_2^2 \leq \phi_{\max} \sum_{l=0}^{\infty} \left\| \boldsymbol{C}_l^{w} \right\|_2^2,$$

by Assumption 4.3. Then, for any $\epsilon > 0$, it follows that

$$\mathbb{P}\left(\frac{\left\| \boldsymbol{W}_{S^c_\pi} \right\|_2}{\sqrt{TM}} \geq K_\epsilon \right) \leq \frac{\sum_{i=1}^{M} \sum_{t=1}^{T} \mathbb{E}\left(w_{i,t}\right)^2}{K_\epsilon^2 TM} \leq \frac{\phi_{\max} \sum_{l=0}^{\infty} \left\| \boldsymbol{C}_l^{w} \right\|_2^2}{K_\epsilon^2} \leq \epsilon,$$

for $K_\epsilon \geq \left(\phi_{\max} \sum_{l=0}^{\infty} \left\| \boldsymbol{C}_l^{w} \right\|_2^2 \right)^{-1/2}$. We use this to further simplify (4.A.22).

First, we show that $\mathscr{B}^c_{w_1,T}$ converges to zero in probability, by noting that

$$\begin{aligned}\mathbb{P}\left(\mathscr{B}^c_{w_1,T}\right) &= \mathbb{P}\left(\left\|\boldsymbol{W}'_{S^c_\pi}\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\lambda_T\omega_{S^c_\pi,\min}}{4} - \frac{\lambda_T\sqrt{s}\left\|\boldsymbol{W}_{S^c_\pi}\right\|_2}{2\sqrt{\phi}T^{1/2-\xi}}\right) \\ &\leq \mathbb{P}\left(\left\|\boldsymbol{W}'_{S^c_\pi}\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\lambda_T\omega_{S^c_\pi,\min}}{4} - \frac{\lambda_T K_\epsilon\sqrt{s}T^{\xi}\sqrt{M}}{2\sqrt{\phi}}\right) + \epsilon.\end{aligned} \tag{4.A.23}$$

It is straightforward to verify that $\{w_{i,t}\epsilon_{y,t}\}$ is a martingale difference sequence. Thus, by application of the Markov bound combined with Burkholder's inequality for martingale difference sequences, it follows that

$$\begin{aligned}\mathbb{P}\left(\left\|\boldsymbol{W}'_{S^c_\pi}\boldsymbol{\epsilon}_y\right\|_2 \geq K_\epsilon\sqrt{TM}\right) &\leq \frac{\sum_{i=1}^M \mathbb{E}\left(\sum_{t=1}^T w_{i,t}\epsilon_{y,t}\right)^2}{K_\epsilon^2 TM} \leq \frac{K\sum_{i=1}^M\sum_{t=1}^T \mathbb{E}\left(w_{i,t}\epsilon_{y,t}\right)^2}{K_\epsilon^2 TM} \\ &\leq \frac{K\sum_{i=1}^M\sum_{t=1}^T\sum_{l_1,l_2=0}^\infty\sum_{j_1,j_2=1}^M \left|c^w_{l_1,i,j_1}\right|\left|c^w_{l_2,i,j_2}\right|\mathbb{E}\left|\epsilon_{j_1,t-1_1}\epsilon_{j_2,t-1_2}\epsilon^2_{y,t}\right|}{K_\epsilon^2 TM} \\ &\leq \frac{K^*\sum_{i=1}^M\left(\sum_{l=0}^\infty\left\|\boldsymbol{c}^w_{l,i}\right\|_1\right)^2}{K_\epsilon^2 M} \leq \frac{K^*\left(\sum_{l=0}^\infty\left\|\boldsymbol{C}^w_l\right\|_1\right)^2}{K_\epsilon} \leq \epsilon,\end{aligned}$$

for $K_\epsilon \geq \left(\frac{K^*\left(\sum_{l=0}^\infty\left\|\boldsymbol{C}^w_l\right\|_1\right)^2}{\epsilon}\right)^{1/2}$. Thus, $\mathbb{P}\left(\mathscr{B}^c_{w_1,T}\right) \to 0$, if

$$\frac{\omega_{S^c_\pi,\min}}{\sqrt{s}T^{\xi}\sqrt{M}} \to \infty, \text{ and } \frac{\lambda_T\omega_{S^c_\delta}}{\sqrt{TM}} \to \infty.$$

These conditions are ensured by Assumption 4.5. Similarly, we bound the probability measure of $\mathscr{B}^c_{w_2,T}$ as defined in (4.A.22) as follows:

$$\begin{aligned}\mathbb{P}\left(\mathscr{B}^c_{w_2,T}\right) &= \mathbb{P}\left(\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\lambda_T\sqrt{\phi}\omega_{S^c_\pi,\min}}{4\left\|\boldsymbol{W}_{S^c_\pi}\right\|_2} - \frac{\lambda_T\sqrt{s}}{2T^{1/2-\xi}}\right) \\ &\leq \mathbb{P}\left(\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 \geq \frac{\lambda_T\sqrt{\phi}\omega_{S^c_\pi,\min}}{4K_\epsilon\sqrt{TM}} - \frac{\lambda_T\sqrt{s}}{2T^{1/2-\xi}}\right) + \epsilon,\end{aligned} \tag{4.A.24}$$

such that, by Lemma 4.1, sufficient conditions for $\mathbb{P}\left(\mathscr{B}^c_{w_2,T}\right) \to 0$ are given by

$$\frac{\omega_{S^c_\pi,\min}}{\sqrt{s}T^{\xi}\sqrt{M}} \to \infty, \text{ and } \frac{\lambda_T\omega_{S^c_\pi,\min}}{(s_\delta \vee \sqrt{s_\pi})\sqrt{TM}} \to \infty.$$

Both of these conditions are satisfied under Assumption 4.5. Hence, $\mathbb{P}\left(\mathscr{B}^c_{w,T}\right) \to 0$, thereby concluding the proof of Theorem 4.1. ■

Remark 4.9. Our approach to bounding $\|\boldsymbol{W}_{S_\pi}\|_2$ does not put any restrictions on the growth rate of M. However, the price for this generality is that $\omega_{S_\pi^c,\min}$ needs to grow by a factor $\sqrt{TM}$ faster to account for the potential divergence of $\|\boldsymbol{W}_{S_\pi}\|_2$. An alternative approach is taken in Medeiros and Mendes (2016), who rely on the Triplex inequality from Jiang (2009) to conclude that $\max_{1\leq i\leq M}\|\boldsymbol{w}_i\|_2 = O_p(1)$. While this approach is less demanding in terms of the initial weights, it puts additional restrictions on the admissible rates of divergence of N and M. Since the growth rates of weights based on a consistent initial consistent estimator are easy to manually adjust (see Section 4.3.2), we proceed without the use of the Triplex inequality.

Proof of Theorem 4.2. First, we recall the definitions $\boldsymbol{V} = (\boldsymbol{V}_1, \boldsymbol{V}_2)$, $\boldsymbol{V}_1 = (\boldsymbol{Z}_{-1,S_\delta}, \boldsymbol{W}_{S_\pi})$, $\boldsymbol{\Omega} = \operatorname{diag}(\boldsymbol{\omega})$ and $\boldsymbol{\Omega}_1 = \operatorname{diag}(\boldsymbol{\omega}_{S_\gamma})$. Since $\hat{\boldsymbol{\gamma}}$ are the minimizers of (4.2.9), they must set the subgradient equation equal to zero:

$$\boldsymbol{V}'(\Delta\boldsymbol{y} - \boldsymbol{V}\hat{\boldsymbol{\gamma}}) - \frac{\lambda_T}{2}\boldsymbol{\Omega}s(\hat{\boldsymbol{\gamma}}) = 0,$$

where $s(\hat{\boldsymbol{\gamma}})$ is the subgradient of $\|\hat{\boldsymbol{\gamma}}\|_1$ (see Hastie et al., 2015, p. 9). Focussing on the first $|S_\gamma|$ equations, we obtain

$$\begin{aligned}&\boldsymbol{V}_1'(\Delta\boldsymbol{y} - \boldsymbol{V}\hat{\boldsymbol{\gamma}}) - \frac{\lambda_T}{2}\boldsymbol{\Omega}_1 s(\hat{\boldsymbol{\gamma}})\\ &\quad= \boldsymbol{V}_1'\left(\boldsymbol{\epsilon}_y - \boldsymbol{V}_1\left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma}\right) - \boldsymbol{V}_2\hat{\boldsymbol{\gamma}}_{S_\gamma^c}\right) - \frac{\lambda_T}{2}\boldsymbol{\Omega}_1 s(\hat{\boldsymbol{\gamma}}) = 0,\end{aligned}$$

from which follows that

$$\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma} = (\boldsymbol{V}_1'\boldsymbol{V}_1)^{-1}\left(\boldsymbol{\epsilon}_y - \boldsymbol{V}_1'\boldsymbol{V}_2\hat{\boldsymbol{\gamma}}_{S_\gamma^c} - \frac{\lambda_T}{2}\boldsymbol{\Omega}_1 s(\hat{\boldsymbol{\gamma}})\right) \tag{4.A.25}$$

Pre-multiplying (4.A.25) by $\boldsymbol{S}_T\boldsymbol{Q}'^{-1}$ and taking the Euclidean norm on both sides, it follows that

$$\begin{aligned}&\left\|\boldsymbol{S}_T\boldsymbol{Q}'^{-1}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma}\right)\right\|_2\\ &\quad\leq \left\|\left(\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{V}_1\boldsymbol{Q}'\boldsymbol{S}_T^{-1}\right)^{-1}\right\|_2\\ &\qquad\times\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\left(\boldsymbol{V}_1'\boldsymbol{\epsilon}_y - \boldsymbol{V}_1'\boldsymbol{V}_2\hat{\boldsymbol{\gamma}}_{S_\gamma^c} - \frac{\lambda_T}{2}\boldsymbol{\Omega}_1 s\left(\hat{\boldsymbol{\gamma}}_{S_\gamma}\right)\right)\right\|_2\\ &\quad\leq \phi^{-1}\left(\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 + \left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{V}_2\hat{\boldsymbol{\gamma}}_{S_\gamma^c}\right\|_2\right.\\ &\qquad\left.+\frac{\lambda_T}{2}\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{\Omega}_1 s\left(\hat{\boldsymbol{\gamma}}_{S_\gamma}\right)\right\|_2\right) + o_p(1),\end{aligned} \tag{4.A.26}$$

by Corollary 4.1. We derive the stochastic order for the three RHS terms of (4.A.26).

First,

$$\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{V}_1'\boldsymbol{\epsilon}_y\right\|_2 = O_p\left(s_\delta \vee \sqrt{s_\pi}\right), \tag{4.A.27}$$

by Lemma 4.1. The second term on the RHS of (4.A.26) vanishes in probability by Theorem 4.1. Finally, the third term is bounded as

$$\begin{aligned}&\frac{\lambda_T}{2}\left\|\boldsymbol{S}_T^{-1}\boldsymbol{Q}\boldsymbol{\Omega}_1\boldsymbol{s}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma}\right)\right\|_2 \leq \frac{\lambda_T}{2}\left\|\boldsymbol{S}_T^{-1}\right\|_2\left\|\boldsymbol{\Omega}_1\boldsymbol{s}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma}\right)\right\|_2 \\ &\leq \frac{\lambda_T\sqrt{s}}{T^{1/2-\xi}} = o\left(s_\delta \vee \sqrt{s_\pi}\right),\end{aligned} \tag{4.A.28}$$

where the last equality follows from part 2 of Assumption 4.5. Hence, plugging (4.A.27)-(4.A.28) into (4.A.26), we conclude that

$$\left\|\boldsymbol{S}_T\boldsymbol{Q}'^{-1}\left(\hat{\boldsymbol{\gamma}}_{S_\gamma} - \boldsymbol{\gamma}_{S_\gamma}\right)\right\|_2 = O_p\left(s_\delta \vee \sqrt{s_\pi}\right),$$

as required. ■

Proof of Theorem 4.3. The analytic expression for the ridge estimator is given by

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_R &= \left(\boldsymbol{V}'\boldsymbol{V} + \lambda_R\boldsymbol{I}_{N+M}\right)^{-1}\boldsymbol{V}'\Delta\boldsymbol{y} \\ &= \left(\boldsymbol{V}'\boldsymbol{V} + \lambda_R\boldsymbol{I}_{N+M}\right)^{-1}\left(\boldsymbol{V}'\boldsymbol{V}\boldsymbol{\gamma} + \boldsymbol{V}'\boldsymbol{\epsilon}_y\right) \\ &= \boldsymbol{\gamma} + \left(\boldsymbol{V}'\boldsymbol{V} + \lambda_R\boldsymbol{I}_{N+M}\right)^{-1}\left(\boldsymbol{V}'\boldsymbol{\epsilon}_y - \lambda_R\boldsymbol{\gamma}\right).\end{aligned} \tag{4.A.29}$$

Let $\boldsymbol{S}_R$ and $\boldsymbol{Q}_R$, after appropriate scaling, (4.A.29) reads as

$$\begin{aligned}\boldsymbol{S}_R\boldsymbol{Q}_R'^{-1}\left(\hat{\boldsymbol{\gamma}}_R - \boldsymbol{\gamma}\right) = &\left(\boldsymbol{S}_R^{-1}\boldsymbol{Q}_R\boldsymbol{V}'\boldsymbol{V}\boldsymbol{Q}_R'\boldsymbol{S}_R^{-1} + \lambda_R\boldsymbol{S}_R^{-2}\right)^{-1} \\ &\times\left(\boldsymbol{S}_R^{-1}\boldsymbol{Q}_R\boldsymbol{V}'\boldsymbol{\epsilon}_y - \lambda_R\boldsymbol{S}_R^{-1}\boldsymbol{Q}_R\boldsymbol{\gamma}\right).\end{aligned} \tag{4.A.30}$$

We proceed by bounding the norms of the three RHS quantities in (4.A.30) as

$$\begin{aligned}\left\|\boldsymbol{S}_R\boldsymbol{Q}_R'^{-1}\left(\hat{\boldsymbol{\gamma}}_R - \boldsymbol{\gamma}\right)\right\|_2 \leq &\left\|\left(\boldsymbol{S}_R^{-1}\boldsymbol{Q}_R\boldsymbol{V}'\boldsymbol{V}\boldsymbol{Q}_R'\boldsymbol{S}_R^{-1} + \lambda_R\boldsymbol{S}_R^{-2}\right)^{-1}\right\|_2 \\ &\times\left(\left\|\boldsymbol{S}_R^{-1}\boldsymbol{Q}_R\boldsymbol{V}'\boldsymbol{\epsilon}_y\right\|_2 + \lambda_R\left\|\boldsymbol{S}_R^{-1}\boldsymbol{Q}_R\boldsymbol{\gamma}\right\|_2\right)\end{aligned} \tag{4.A.31}$$

Focussing on the first RHS term of (4.A.31), we note that

$$\begin{aligned}&\left\|\left(\boldsymbol{S}_R^{-1}\boldsymbol{Q}_R\boldsymbol{V}'\boldsymbol{V}\boldsymbol{Q}_R'\boldsymbol{S}_R^{-1} + \lambda_R\boldsymbol{S}_R^{-2}\right)^{-1}\right\|_2 \\ &\quad\geq \frac{1}{\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_R\right) + \frac{\lambda_R}{T^2}} + o_p(1) \geq \frac{1}{\phi_R} + o_p(1),\end{aligned}$$

by part 1 of Corollary 4.2. The stochastic order of the second RHS term is given by part 2 of Corollary 4.2 as

$$\left\|\boldsymbol{S}_R^{-1}\boldsymbol{Q}_R\boldsymbol{V}'\boldsymbol{\epsilon}_y\right\|_2 = O_p\left(N_\delta \vee \sqrt{M_\pi}\right).$$

The third and final RHS is deterministically bounded by

$$\begin{aligned}\lambda_R\left\|\boldsymbol{S}_R^{-1}\boldsymbol{Q}_R\boldsymbol{\gamma}\right\|_2 &\leq \frac{\lambda_R}{\sqrt{T}}\left(\left\|\left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1/2}\boldsymbol{B}'\boldsymbol{\delta}\right\|_2 + \|\boldsymbol{\pi}\|_2\right)\\ &\leq \frac{\lambda_R}{\sqrt{T}}\left(\sqrt{|S_\delta|}\,\|\boldsymbol{\delta}\|_\infty + \sqrt{|S_\pi|}\,\|\boldsymbol{\pi}\|_\infty\right) = O\left(\frac{\lambda_R\left(\sqrt{|S_\delta|}\vee\sqrt{|S_\pi|}\right)}{\sqrt{T}}\right).\end{aligned}$$

As a result, we obtain the stochastic order of (4.A.30) as

$$\begin{aligned}\left\|\boldsymbol{S}_R\boldsymbol{Q}_R'^{-1}\left(\hat{\gamma}_R-\boldsymbol{\gamma}\right)\right\|_2 &= O_p\left(N_\delta\vee\sqrt{M_\pi}\right) + O_p\left(\frac{\lambda_R\left(\sqrt{|S_\delta|}\vee\sqrt{|S_\pi|}\right)}{\sqrt{T}}\right)\\ &= O_p\left(N_\delta\vee\sqrt{M_\pi}\right)\end{aligned}$$

where the last equality follows from the assumption that $\lambda_R \leq \frac{K_R\left(N_\delta\vee\sqrt{M_\pi}\right)\sqrt{T}}{\left(\sqrt{|S_\delta|}\vee\sqrt{|S_\pi|}\right)}$. ■

4.A.3 Satisfying Assumption 4.4

Sample covariance matrices appear on several instances in the sets described in Proposition 4.1. Bounding appropriate norms of (the inverses of) these matrices turns out to be crucial in the proofs of our main results. One of the norms that has attractive theoretical properties is the spectral norm, which, when applied to the inverse of a symmetric positive definite matrix, can be bounded with the use of a lower bound on the minimum eigenvalue, thereby motivating the use of Assumption 4.4. However, the feasibility of such minimum eigenvalue bounds is difficult to verify directly on general sample covariance matrices. One method of verification for the positive lower bound on the minimum eigenvalue of $\hat{\boldsymbol{\Sigma}}_{11}$, i.e. part 1 of Assumption 4.4, is by restricting the eigenvalues of a simpler approximating matrix. The behaviour of the eigenvalues of this approximating matrix can be shown to carry over to the sample covariance matrix, based on either of the following two results.

Lemma 4.3. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ denote two s-dimensional square non-negative definite matrices. Then,*

(i) for all $i = 1, \ldots, s$*, it holds that*

$$|\lambda_i(\boldsymbol{A}) - \lambda_i(\boldsymbol{B})| \leq \|\boldsymbol{A} - \boldsymbol{B}\|_2,$$

(ii) if $\|\boldsymbol{A} - \boldsymbol{B}\|_{\max} \leq \delta$*, then* $\lambda_{\min}(\boldsymbol{B}) \geq \lambda_{\min}(\boldsymbol{A}) - s\delta$*.*

Proof of Lemma 4.3. **Proof of Part (i)**
This is a well-known consequence of the additive Weyl inequalities, see for example Horn et al. (1994, Theorem 3.3.16). Note in particular, that when $\boldsymbol{A}$ is a symmetric positive definite matrix, it holds that $\sigma_i(\boldsymbol{A}) = \lambda_i(\boldsymbol{A})$, i.e. the singular values are equal to the eigenvalues.

Proof of Part (ii)
This corresponds to Lemma 6.17 in Bühlmann and Van De Geer (2011) and is described in similar form in Lemma 3 in Medeiros and Mendes (2016). For the sake of completion, we repeat the short proof here. Take $\boldsymbol{x} \in \mathbb{R}^s \setminus \{0\}$. Then,

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x} \leq |\boldsymbol{x}'(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{x}| \leq \|\boldsymbol{x}\|_1 |(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{x}|_\infty \leq \|\boldsymbol{x}\|_1^2 \delta \leq \boldsymbol{x}'\boldsymbol{x} s\delta,$$

from which clearly follows that

$$\frac{x'\boldsymbol{B}x}{x'x} \geq \frac{x'\boldsymbol{A}x}{x'x} - s\delta.$$

Taking the infimum on both sides completes the proof. ■

An important consequence of Lemma 4.3 is that a bound on the minimum eigenvalue of $\boldsymbol{A}$, automatically results in a bound for the minimum eigenvalue of $\boldsymbol{B}$, the latter depending on the maximum distance between the elements of the two matrices. We use this to derive the following result.

Lemma 4.4. *Define* $\boldsymbol{\Sigma}_{11} = \mathbb{E}\left(\boldsymbol{v}_{1,t}\boldsymbol{v}_{1,t}'\right)$*. Furthermore, assume that* $\frac{s_\pi}{\sqrt{T}} \to 0$ *and* $\lambda_{\min}(\boldsymbol{\Sigma}_{11}) \geq 2\phi$ *for some* $\phi > 0$*. Then, under Assumptions 4.1-4.3,*

$$\mathbb{P}\left(\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{11}\right) \geq \phi\right) \to 1$$

as $T, s_\delta, s_\pi \to \infty$*.*

Proof of Lemma 4.4. We prove Lemma 4.4 by showing that $\left\|\hat{\boldsymbol{\Sigma}}_{11} - \boldsymbol{\Sigma}_{11}\right\|_2 \xrightarrow{p} 0$ as $T, s_\pi \to \infty$, after which application of part (i) of Lemma 4.3 leads to the desired result. Chen et al. (2013) derive the convergence rates for thresholded estimates of high-dimensional covariance matrices, based on the functional dependence measure

in Wu (2005). The key feature of this dependence measure is the construction of a coupled version of the stochastic process, which in our setting results in the process $\boldsymbol{v}_{1,t}^* = \sum_{l=0}^{\infty} \boldsymbol{C}_l^v \boldsymbol{\epsilon}_{t-l}^*$, where $\boldsymbol{\epsilon}_t^* = \boldsymbol{\epsilon}_t$ for $t \neq 0$ and $\boldsymbol{\epsilon}_0^*$ is an i.i.d. copy of $\boldsymbol{\epsilon}_0$. By Assumption 4.1, for any $w \leq 2m$, the functional dependence measure for element $v_{1,j,t}$ is bounded by

$$\begin{aligned}\theta_{j,t,w} &= \left\|v_{1,j,t} - v_{1,j,t}^*\right\|_w = \left\|\sum_{i=1}^{N}\sum_{l=0}^{\infty} c_{l,j,i}^v \left(\epsilon_{i,t-l} - \epsilon_{i,t-l}^*\right)\right\|_w = \left\|\sum_{i=1}^{N} c_{t,j,i}^v \left(\epsilon_{i,0} - \epsilon_{i,0}^*\right)\right\|_w \\ &\leq \sum_{i=1}^{N} \left|c_{t,j,i}^v\right| \left\|\epsilon_{i,0} - \epsilon_{i,0}^*\right\|_w \leq \left\|\epsilon_{i,0} - \epsilon_{i,0}^*\right\|_w \left\|\boldsymbol{c}_{t,j}^v\right\|_1 \leq K \left\|\boldsymbol{c}_{t,j}^v\right\|_1,\end{aligned}$$

where $\boldsymbol{c}_{t,j}^v$ is the j-th row of $\boldsymbol{C}_t^v$ and $\|\cdot\|_w = (\mathbb{E}(\cdot)^w)^{1/w}$. Then, with the addition of Assumption 4.3, it holds that

$$\Theta_{k,w} = \max_{1\leq j\leq N} \sum_{l=k}^{\infty} \theta_{j,l,w} \leq K \sum_{l=k}^{\infty} \|\boldsymbol{C}_l^v\|_\infty = O\left(k^{-1}\right),$$

for all $k > 0$. Therefore, the conditions in Theorem 2.1 of Chen et al. (2013) are satisfied. From this theorem, it follows that $\mathbb{E}\left\|\hat{\boldsymbol{\Sigma}}_{11} - \boldsymbol{\Sigma}_{11}\right\|_F^2 = O\left(\frac{s_\pi^2}{T}\right)$, by taking the limit of the threshold value as it approaches zero. Thus, application of the Markov inequality shows that $\left\|\hat{\boldsymbol{\Sigma}}_{11} - \boldsymbol{\Sigma}_{11}\right\|_2 \xrightarrow{p} 0$ as $T, s_\pi \to \infty$ with $\frac{s_\pi}{\sqrt{T}} \to 0$. By part (i) of Lemma 4.3, this implies that $\left|\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{11}\right) - \lambda_{\min}\left(\boldsymbol{\Sigma}_{11}\right)\right| \xrightarrow{p} 0$. Then,

$$\begin{aligned}\mathbb{P}\left(\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{11}\right) \geq \phi\right) &= \mathbb{P}\left(\left|\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{11}\right) - \lambda_{\min}\left(\boldsymbol{\Sigma}_{11}\right) + \lambda_{\min}\left(\boldsymbol{\Sigma}_{11}\right)\right| \geq \phi\right) \\ &\geq \mathbb{P}\left(\lambda_{\min}\left(\boldsymbol{\Sigma}_{11}\right) - \left|\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{11}\right) - \lambda_{\min}\left(\boldsymbol{\Sigma}_{11}\right)\right| \geq \phi\right) \\ &= \mathbb{P}\left(\left|\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{11}\right) - \lambda_{\min}\left(\boldsymbol{\Sigma}_{11}\right)\right| \leq \lambda_{\min}\left(\boldsymbol{\Sigma}_{11}\right) - \phi\right) \\ &\geq \mathbb{P}\left(\left|\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{11}\right) - \lambda_{\min}\left(\boldsymbol{\Sigma}_{11}\right)\right| \leq \phi\right) \to 1,\end{aligned}$$

as required. ■

Next we focus on bounding the minimum eigenvalue of

$$\hat{\boldsymbol{\Sigma}}_{22} = \frac{s_\delta}{T^2}\sum_{t=1}^{T} \boldsymbol{v}_{2,t}\boldsymbol{v}_{2,t}' = \frac{s_\delta}{T^2}\boldsymbol{B}_{S_\delta,\perp}'\left(\sum_{t=1}^{T} \boldsymbol{z}_{S_\delta,t}\boldsymbol{z}_{S_\delta,t}'\right)\boldsymbol{B}_{S_\delta,\perp}$$

Contrary to $\hat{\boldsymbol{\Sigma}}_{11}$, the matrix $\hat{\boldsymbol{\Sigma}}_{22}$ does not converge in probability to a deterministic matrix. The following result is used in Remark 4.2 and demonstrates the issues with

the collinearity inducing property of integrated variables in high dimensions.

Lemma 4.5. *Define an s-dimensional white noise sequence* $\boldsymbol{u}_t \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_s)$ *and let* $\boldsymbol{h}_t = \sum_{j=1}^{t} \boldsymbol{u}_j$. *Then, as* $s, T \to \infty$,

$$\mathbb{P}\left(\lambda_{\min}\left(\frac{1}{T^2}\sum_{t=1}^{T}\boldsymbol{h}_t\boldsymbol{h}_t'\right) > \phi\right) \to 0, \tag{4.A.32}$$

for any $\phi > 0$.

Proof. We show that $\lambda_{\min}\left(\frac{1}{T^2}\sum_{t=1}^{T}\boldsymbol{h}_t\boldsymbol{h}_t'\right) \xrightarrow{p} 0$, as $T, s \to \infty$. Let $E = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_s\}$ be the collection of basis vectors. Since $E \subset \mathbb{R}^s$, we have for any $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\lambda_{\min}\left(\frac{1}{T^2}\sum_{t=1}^{T}\boldsymbol{h}_t\boldsymbol{h}_t'\right) > \phi\right) \leq \mathbb{P}\left(\min_{\boldsymbol{x}\in E}\boldsymbol{x}'\left(\lambda_{\min}\left(\frac{1}{T^2}\sum_{t=1}^{T}\boldsymbol{h}_t\boldsymbol{h}_t'\right)\right)\boldsymbol{x} > \phi\right) \\
&= \mathbb{P}\left(\min_{1\leq i\leq s}\frac{1}{T^2}\sum_{t=1}^{T}h_{i,t}^2 > \phi\right) = \mathbb{P}\left(\frac{1}{T^2}\sum_{t=1}^{T}h_{1,t}^2 > \phi\right)^s \\
&\leq \left\{\mathbb{P}\left(\int_0^1 W^2(r)dr > \phi\right) + \left|\mathbb{P}\left(\frac{1}{T^2}\sum_{t=1}^{T}h_{1,t}^2 > \phi\right) - \mathbb{P}\left(\int_0^1 W^2(r)dr > \phi\right)\right|\right\}^s,
\end{aligned}$$

where $W(r)$ is a standard univariate Brownian Motion. First, assume that

$$\mathbb{P}\left(\int_0^1 W^2(r)dr > \phi\right) \leq 1 - 2\epsilon(\phi), \tag{4.A.33}$$

for some $\epsilon(\phi) > 0$. Then, by the functional central limit theorem, there exists a T^* such that

$$\left|\mathbb{P}\left(\frac{1}{T^2}\sum_{t=1}^{T}h_{1,t}^2 > \phi\right) - \mathbb{P}\left(\int_0^1 W^2(r)dr > \phi\right)\right| \leq \epsilon(\phi)$$

for all $T > T^*$. Consequently, for large enough T,

$$\mathbb{P}\left(\lambda_{\min}\left(\frac{1}{T^2}\sum_{t=1}^{T}\boldsymbol{h}_t\boldsymbol{h}_t'\right) > \phi\right) \leq (1 - \epsilon(\phi))^s \to 0$$

as $s, T \to \infty$, which is the claim of Lemma 4.5. Hence, all that is left is to verify the truth of (4.A.33).

First, note that

$$\mathbb{P}\left(\int_0^1 W^2(r)dr \geq \phi\right) \leq \mathbb{P}\left(\sup_{0\leq r\leq 1}|W(r)| \geq \sqrt{\phi}\right) = 1-\mathbb{P}\left(\sup_{0\leq r\leq 1}|W(r)| < \sqrt{\phi}\right).$$

We show that for every $\phi > 0$, it holds that $\mathbb{P}\left(\sup_{0\leq r\leq 1}|W(r)| < \sqrt{\phi}\right) \geq 2\epsilon(\phi) > 0$ for some $\epsilon(\phi) > 0$. Let W_1 and W_2 denote two independent standard Brownian motions over the interval $[0,1]$ and note that we may construct an additional standard Brownian motion as $W_\Delta = (W_1 - W_2)/\sqrt{2}$. The sample paths of W_1, W_2, W_Δ lie in the function space $C([0,1])$, which contains all continuous functions from the unit interval to $\mathbb{R}$ and is equipped with the supremum norm $\|f\|_\infty = \max\{|f(x)| \mid x \in [0,1]\}$. It is well-known that $C([0,1])$ is a separable metric space (e.g. Davidson, 1994, p. 438). Then, define $B(y, \sqrt{2\phi}) = \{x \in C([0,1]) \mid \|x-y\|_\infty \leq \sqrt{2\phi}\}$ and note that by Theorem 5.6 of Davidson (1994) there exists a countable collection of elements $\{x_1, x_2, \ldots\} \subset C([0,1])$ such that $C([0,1]) \subseteq \bigcup_i B(x_i, \sqrt{2\phi})$. By countable additivity,

$$1 = \mathbb{P}\left(W_1 \in C([0,1])\right) = \mathbb{P}\left(W_1 \in \bigcup_i B(x_i, \sqrt{2\phi})\right) \leq \sum_i \mathbb{P}\left(W_1 \in B(x_i, \sqrt{2\phi})\right). \tag{4.A.34}$$

Furthermore, it must be true that there exists a $B(x_i, \sqrt{2\phi})$ with $\mathbb{P}\left(W_1 \in B(x_i, \sqrt{2\phi})\right) = q > 0$, because otherwise the RHS of (4.A.34) would be zero, resulting in a contradiction. Since W_1 and W_2 are independent, we conclude that

$$\mathbb{P}\left(\sup_{0\leq r\leq 1}|W_\Delta(r)| < \sqrt{\phi}\right) \geq \mathbb{P}\left(W_1 \in B(x_i, \sqrt{2\phi}), W_2 \in B(x_i, \sqrt{2\phi})\right) = q^2 > 0.$$

Since q depends only on ϕ, we may write $2\epsilon(\phi) = q^2$, thereby completing the proof. ■

Hence, we aim to bound $\hat{\boldsymbol{\Sigma}}_{22}$, which contains a scaling by $\frac{s_\delta}{T^2}$, under varying additional assumptions on the DGP and the growth rate of s_δ. The first bound that we derive assumes normality of the errors and requires $\frac{s_\delta}{T^{1/2}} \to 0$.

Lemma 4.6. *Let $\hat{\boldsymbol{\Sigma}}_{22}$ be as defined in Assumption 4.4 and assume that $\boldsymbol{\epsilon}_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$. Then, under Assumptions 4.1-4.3, there exists a constant $\zeta > 0$ such that*

$$\mathbb{P}\left(\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{22}\right) \geq \zeta\right) \to 1,$$

as $s_\delta, T \to \infty$ *with* $\frac{s_\delta}{T^{1/2}} \to 0$.

Proof. First, letting $\boldsymbol{u}_t = \boldsymbol{C}_{S_\delta}(L)\boldsymbol{\epsilon}_t$, we decompose $\hat{\boldsymbol{\Sigma}}_{22}$ into

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_{22} &= \boldsymbol{B}'_{S_\delta,\perp}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{z}_{S_\delta,t}\boldsymbol{z}'_{S_\delta,t}\right)\boldsymbol{B}_{S_\delta,\perp} \\
&= \boldsymbol{B}'_{S_\delta,\perp}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\left(\boldsymbol{C}_{S_\delta}\boldsymbol{s}_{t-1}+\boldsymbol{u}_{t-1}\right)\left(\boldsymbol{C}_{S_\delta}\boldsymbol{s}_{t-1}+\boldsymbol{u}_{t-1}\right)'\right)\boldsymbol{B}_{S_\delta,\perp} \\
&= \boldsymbol{B}'_{S_\delta,\perp}\boldsymbol{C}_{S_\delta}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{s}_{t-1}\boldsymbol{s}'_{t-1}\right)\boldsymbol{C}'_{S_\delta}\boldsymbol{B}_{S_\delta,\perp} \\
&\quad + \boldsymbol{B}'_{S_\delta,\perp}\boldsymbol{C}_{S_\delta}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{s}_{t-1}\boldsymbol{u}'_{t-1}\right)\boldsymbol{B}_{S_\delta,\perp} \\
&\quad + \boldsymbol{B}'_{S_\delta,\perp}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{u}_{t-1}\boldsymbol{s}'_{t-1}\right)\boldsymbol{C}'_{S_\delta}\boldsymbol{B}_{S_\delta,\perp} \\
&\quad + \boldsymbol{B}'_{S_\delta,\perp}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{u}_{t-1}\boldsymbol{u}'_{t-1}\right)\boldsymbol{B}_{S_\delta,\perp} \\
&=: A_1 + A_2 + A_2' + A_3,
\end{aligned} \tag{4.A.35}$$

such that

$$\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{22}\right) \geq \lambda_{\min}(A_1) - 2\left\|\boldsymbol{A}_2\right\|_2 - \left\|\boldsymbol{A}_3\right\|_2. \tag{4.A.36}$$

We show that there exists a $\zeta > 0$ such that

$$\mathbb{P}\left(\lambda_{\min}\left(\boldsymbol{A}_1\right) > \zeta\right) \to 1,$$

whereas

$$\mathbb{P}\left(\left\|\boldsymbol{A}_2\right\|_2 > \zeta\right) \to 0 \text{ and } \mathbb{P}\left(\left\|\boldsymbol{A}_3\right\|_2 > \zeta\right) \to 0$$

as $s_\delta, T \to \infty$.

By the assumption that $\boldsymbol{\epsilon}_t \overset{i.i.d.}{\sim} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon\right)$, it holds that

$$\begin{aligned}
\boldsymbol{A}_1 &= \boldsymbol{B}'_{S_\delta,\perp}\boldsymbol{C}_{S_\delta}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{s}_{t-1}\boldsymbol{s}'_{t-1}\right)\boldsymbol{C}'_{S_\delta}\boldsymbol{B}_{S_\delta,\perp} \\
&\overset{d}{=} \boldsymbol{R}^{1/2}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{s}_{G,t-1}\boldsymbol{s}'_{G,t-1}\right)\boldsymbol{R}^{1/2},
\end{aligned}$$

where $\boldsymbol{R} = \boldsymbol{B}'_{S_\delta,\perp}\boldsymbol{C}_{S_\delta}\boldsymbol{\Sigma}_\epsilon\boldsymbol{C}'_{S_\delta}\boldsymbol{B}_{S_\delta,\perp}$, and $\boldsymbol{s}_{G,t} = \sum_{s=1}^{t}\boldsymbol{\epsilon}_{G,s}$ with $\boldsymbol{\epsilon}_{G,s} \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{s_\delta})$ being an s_δ-dimensional Gaussian white noise process. Furthermore, for any $\boldsymbol{x} \in \mathbb{R}^{s_\delta}$ with $\boldsymbol{x}'\boldsymbol{x} = 1$,

$$\boldsymbol{x}'\boldsymbol{B}'_{S_\delta,\perp}\boldsymbol{C}_{S_\delta}\boldsymbol{\Sigma}_\epsilon\boldsymbol{C}'_{S_\delta}\boldsymbol{B}_{S_\delta,\perp}\boldsymbol{x} = \boldsymbol{y}'\boldsymbol{\Sigma}_\epsilon\boldsymbol{y} \geq \phi_{\min},$$

where $\boldsymbol{y} \neq \boldsymbol{0}$, because $\boldsymbol{C}'_{S_\delta}\boldsymbol{B}_{S_\delta,\perp}$ has full column rank by construction, and the inequality follows by Assumption 4.1. Thus,

$$\begin{aligned}\lambda_{\min}(\boldsymbol{A}_1) &\geq \lambda_{\min}(\boldsymbol{R})\,\lambda_{\min}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{s}_{G,t-1}\boldsymbol{s}'_{G,t-1}\right)\\ &\geq \phi_{\min}\lambda_{\min}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{s}_{G,t-1}\boldsymbol{s}'_{G,t-1}\right).\end{aligned}$$

Let $\boldsymbol{S}_G = (\boldsymbol{s}_{G,1}, \ldots, \boldsymbol{s}_{G,T})'$ and $\boldsymbol{E}_G = (\boldsymbol{\epsilon}_{G,1}, \ldots, \boldsymbol{\epsilon}_{G,T})'$. Note that we can rewrite $\boldsymbol{S}_G = \boldsymbol{U}\boldsymbol{E}_G$, where $\boldsymbol{U}$ is a lower triangular matrix with ones on and below the diagonal. Furthermore, we can decompose $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'$, where $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_T)$. By Lemma 1 in Akesson and Lehoczky (1998), it holds that

$$\lambda_t^{-1} = 4\sin^2\left(\frac{\omega_t}{2}\right) = 2(1 - \cos\omega_t), \tag{4.A.37}$$

with $\omega_t = \frac{(2t-1)\pi}{2T+1}$. The second equality in (4.A.37) is based on the identity $\cos(2\alpha) = 1 - 2\sin^2(\alpha)$. It follows that

$$\boldsymbol{S}'_G\boldsymbol{S}_G = \boldsymbol{E}'_G\boldsymbol{U}'\boldsymbol{U}\boldsymbol{E}_G = \boldsymbol{E}'_G\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'\boldsymbol{E}_G = \tilde{\boldsymbol{E}}\boldsymbol{\Lambda}\tilde{\boldsymbol{E}},$$

where $\tilde{\boldsymbol{E}} = \boldsymbol{V}'\boldsymbol{E}_G$. Note that, as a result of the rotational invariance of the multivariate normal distribution, $\tilde{\boldsymbol{E}}$ is again an $(T \times s_\delta)$-dimensional matrix with independent standard normal entries. Define $\mathbb{R}_G = \{\boldsymbol{x} \in \mathbb{R}^{s_\delta} : \boldsymbol{x}'\boldsymbol{x} = 1\}$. Let $\boldsymbol{y}_x = \boldsymbol{V}'\boldsymbol{E}_G\boldsymbol{x}$. Then, following a similar strategy as in the proof of Remark 3.5 in Zhang et al.

(2019a),

$$\begin{aligned}
&\lambda_{\min}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{s}_{G,t-1}\boldsymbol{s}_{G,t-1}'\right) = \frac{s_\delta}{T^2}\lambda_{\min}\left(\boldsymbol{E}_G'\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'\boldsymbol{E}_G\right), \\
&= \frac{s_\delta}{T^2}\min_{\boldsymbol{x}\in R_G}\boldsymbol{x}'\boldsymbol{E}_G'\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'\boldsymbol{E}_G\boldsymbol{x} = \frac{s_\delta}{T}\min_{\boldsymbol{x}\in R_G}\frac{\boldsymbol{y}_x'\boldsymbol{y}_x}{T}\frac{\boldsymbol{y}_x'\boldsymbol{\Lambda}\boldsymbol{y}_x}{\boldsymbol{y}_x'\boldsymbol{y}_x} \\
&\geq \frac{s_\delta}{T}\left(\min_{\boldsymbol{x}\in R_G}\frac{\boldsymbol{y}_x'\boldsymbol{y}_x}{T}\right)\left(\min_{\boldsymbol{x}\in R_G}\frac{\boldsymbol{y}_x'\boldsymbol{\Lambda}\boldsymbol{y}_x}{\boldsymbol{y}_x'\boldsymbol{y}_x}\right) \\
&\geq \frac{s_\delta}{T}\lambda_{\min}\left(\frac{\tilde{\boldsymbol{E}}'\tilde{\boldsymbol{E}}}{T}\right)\left(\min_{\boldsymbol{x}\in R_G}\frac{\sum_{j=1}^{s_\delta}y_{x,j}^2\lambda_j}{\boldsymbol{y}_x'\boldsymbol{y}_x}\right) \\
&\geq \frac{ks_\delta}{T^2}\frac{\lambda_{\min}\left(\tilde{\boldsymbol{E}}'\tilde{\boldsymbol{E}}/T\right)}{\lambda_{\max}\left(\tilde{\boldsymbol{E}}'\tilde{\boldsymbol{E}}/T\right)}\min_{\boldsymbol{x}\in R_G}\frac{1}{k}\sum_{j=1}^{k}y_{x,j}^2\lambda_{k+1}.
\end{aligned} \tag{4.A.38}$$

Furthermore, it holds that

$$\lambda_{k+1} = \frac{1}{2(1-\cos\omega_{k+1})} \geq \frac{1}{2\omega_{k+1}^2} = \frac{(2T+1)^2}{8\pi^2k^2} \geq \frac{T^2}{2\pi^2k^2}. \tag{4.A.39}$$

In addition, by Theorem 2.1 in Chen et al. (2013) it follows that, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left\|\frac{\tilde{\boldsymbol{E}}'\tilde{\boldsymbol{E}}}{T} - \boldsymbol{I}_{s_\delta}\right\|_2 > \epsilon\right) \to 0$$

as $s_\delta, T \to \infty$ with $\frac{s_\delta}{T^{1/2}} \to 0$. As a consequence of Weyl's inequality in Lemma 4.3, this implies that

$$\mathbb{P}\left(\left|\frac{\lambda_{\min}\left(\tilde{\boldsymbol{E}}'\tilde{\boldsymbol{E}}/T\right)}{\lambda_{\max}\left(\tilde{\boldsymbol{E}}'\tilde{\boldsymbol{E}}/T\right)} - 1\right| > \epsilon\right) \to 0, \tag{4.A.40}$$

as $s_\delta, T \to \infty$ with $\frac{s_\delta}{T^{1/2}} \to 0$. Finally, define $\boldsymbol{V}^k = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$, and note that $\left(\boldsymbol{V}^k\right)'\tilde{\boldsymbol{E}}$ is a $(k \times s_\delta)$-dimensional matrix with independent standard normal entries. Then, choosing k such that $\frac{s_\delta}{k} \to y \in (0,1)$, it follows from Theorem 1 in Bai and Yin (1993) that

$$\lim\lambda_{\min}\left(\frac{\boldsymbol{E}_G'\boldsymbol{V}^K\left(\boldsymbol{V}^k\right)'\boldsymbol{E}_G}{k}\right) = (1-\sqrt{y})^2, \text{ a.s.} \tag{4.A.41}$$

Then, plugging (4.A.39)-(4.A.41) into (4.A.38), we obtain

$$\begin{aligned}\lambda_{\min}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{s}_{G,t-1}\boldsymbol{s}_{G,t-1}'\right) &\geq \frac{s_\delta}{2\pi^2 k}\frac{\lambda_{\min}\left(\tilde{\boldsymbol{E}}'\tilde{\boldsymbol{E}}/T\right)}{\lambda_{\max}\left(\tilde{\boldsymbol{E}}'\tilde{\boldsymbol{E}}/T\right)}\min_{\boldsymbol{x}\in R_G}\frac{1}{k}\sum_{j=1}^{k}y_{x,j}^2 \\ &= \frac{s_\delta}{2\pi^2 k}\frac{\lambda_{\min}\left(\tilde{\boldsymbol{E}}'\tilde{\boldsymbol{E}}/T\right)}{\lambda_{\max}\left(\tilde{\boldsymbol{E}}'\tilde{\boldsymbol{E}}/T\right)}\lambda_{\min}\left(\frac{\tilde{\boldsymbol{E}}'\boldsymbol{V}^K\left(\boldsymbol{V}^k\right)'\tilde{\boldsymbol{E}}}{k}\right) \to \frac{y(1-\sqrt{y})^2}{2\pi^2},\end{aligned} \tag{4.A.42}$$

in probability as $s_\delta, T \to \infty$ with $\frac{s_\delta}{T^{1/2}} \to 0$.

It remains to show that $\|\boldsymbol{A}_2\|_2$ and $\|\boldsymbol{A}_3\|_2$ converge in probability to zero as $s_\delta, T \to \infty$. First, applying the Beveridge-Nelson decomposition to $\boldsymbol{C}_{S_\delta}(L) = \boldsymbol{C}_{S_\delta}(1) + \boldsymbol{C}^*_{S_\delta}(L)(1-L)$, and letting $\boldsymbol{\eta}_t = \boldsymbol{C}^*_{S_\delta}(L)\boldsymbol{\epsilon}_t$, we can rewrite

$$\begin{aligned}&\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{s}_{t-1}\boldsymbol{u}_{t-1}'\right)\boldsymbol{B}_{S_\delta,\perp} \\ &\quad= \boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{s}_{t-2}\boldsymbol{\epsilon}_{t-1}'\right)\boldsymbol{C}_{S_\delta}(1)\boldsymbol{B}_{S_\delta,\perp} \\ &\qquad+ \boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{\epsilon}_{t-1}\boldsymbol{\epsilon}_{t-1}'\right)\boldsymbol{C}_{S_\delta}(1)\boldsymbol{B}_{S_\delta,\perp} \\ &\qquad+ \boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{s}_{t-1}\left(\boldsymbol{\eta}_{t-1}-\boldsymbol{\eta}_{t-2}\right)'\right)\boldsymbol{B}_{S_\delta,\perp}.\end{aligned} \tag{4.A.43}$$

Furthermore, using summation by parts, we can further simplify the last term on the RHS of (4.A.43) to

$$\begin{aligned}&\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{s}_{t-1}\left(\boldsymbol{\eta}_{t-1}-\boldsymbol{\eta}_{t-2}\right)'\right)\boldsymbol{B}_{S_\delta,\perp} \\ &\quad= \boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\left(\frac{s_\delta}{T^2}\boldsymbol{s}_{T-1}\boldsymbol{\eta}_{T-1}'\right)\boldsymbol{B}_{S_\delta,\perp} \\ &\qquad- \boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T-1}\boldsymbol{\epsilon}_t\boldsymbol{\eta}_{t-1}'\right)\boldsymbol{B}_{S_\delta,\perp}.\end{aligned} \tag{4.A.44}$$

Then, plugging (4.A.44) into (4.A.43), we obtain the lengthy expression

$$\begin{aligned}
&\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{s}_{t-1}\boldsymbol{u}_{t-1}'\right)\boldsymbol{B}_{S_\delta,\perp}\\
&\quad=\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{s}_{t-2}\boldsymbol{\epsilon}_{t-1}'\right)\boldsymbol{C}_{S_\delta}(1)\boldsymbol{B}_{S_\delta,\perp}\\
&\qquad+\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{\epsilon}_{t-1}\boldsymbol{\epsilon}_{t-1}'\right)\boldsymbol{C}_{S_\delta}(1)\boldsymbol{B}_{S_\delta,\perp}\\
&\qquad+\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\left(\frac{s_\delta}{T^2}\boldsymbol{s}_{T-1}\boldsymbol{\eta}_{T-1}'\right)\boldsymbol{B}_{S_\delta,\perp}\\
&\qquad-\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T-1}\boldsymbol{\epsilon}_{t}\boldsymbol{\eta}_{t-1}'\right)\boldsymbol{B}_{S_\delta,\perp}=:\sum_{i=1}^{4}\boldsymbol{D}_i,
\end{aligned}\tag{4.A.45}$$

with each $\boldsymbol{D}_i$ corresponding to the i-th term on the RHS of the first equation. Thus, we may derive the convergence rate of $\boldsymbol{A}_2$ in (4.A.35) based on the rates of the individual terms $\boldsymbol{D}_i$ in (4.A.45). For notational convenience, let $\boldsymbol{a}_i=\boldsymbol{C}_{S_\delta}'\boldsymbol{\beta}_{S_\delta,\perp,i}$ and $\boldsymbol{b}_j=\boldsymbol{C}_{S_\delta}(1)\boldsymbol{\beta}_{S_\delta,\perp,j}$. Starting with the first term, we obtain

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{D}_1\|_2\geq\zeta\right)&\leq\mathbb{P}\left(\sum_{i,j=1}^{s_\delta}\left(\frac{s_\delta}{T^2}\sum_{t=3}^{T}\boldsymbol{a}_i'\boldsymbol{s}_{t-2}\boldsymbol{\epsilon}_{t-1}'\boldsymbol{b}_j\right)^2\geq\zeta^2\right)\\
&\leq\frac{s_\delta^2\sum_{i,j=1}^{s_\delta}\mathbb{E}\left(\sum_{t=3}^{T}\boldsymbol{a}_i'\boldsymbol{s}_{t-2}\boldsymbol{\epsilon}_{t-1}'\boldsymbol{b}_j\right)^2}{T^4\zeta^2}.
\end{aligned}\tag{4.A.46}$$

Then, noting that $\{\boldsymbol{a}_i'\boldsymbol{s}_{t-2}\boldsymbol{\epsilon}_{t-1}'\boldsymbol{b}_j\}$ is a martingale difference sequence, it follows from Burkholder's inequality in combination with the C_r-inequality that we can bound the expectation by

$$\begin{aligned}
\mathbb{E}\left(\sum_{t=3}^{T}\boldsymbol{a}_i'\boldsymbol{s}_{t-2}\boldsymbol{\epsilon}_{t-1}'\boldsymbol{b}_j\right)^2&\leq K\sum_{t=3}^{T}\mathbb{E}\left(\boldsymbol{a}_i\boldsymbol{s}_{t-2}\boldsymbol{\epsilon}_{t-1}'\boldsymbol{b}_j\right)^2\\
&=K\sum_{t=3}^{T}\mathbb{E}\left(\boldsymbol{a}_i'\boldsymbol{s}_{t-2}\right)^2\mathbb{E}\left(\boldsymbol{b}_j'\boldsymbol{\epsilon}_{t-1}\right)^2=K\left(\boldsymbol{b}_j'\boldsymbol{\Sigma}_\epsilon\boldsymbol{b}_j\right)\left(\boldsymbol{a}_i'\boldsymbol{\Sigma}_\epsilon\boldsymbol{a}_i\right)\sum_{t=1}^{T-2}t\\
&\leq T^2K\left\|\boldsymbol{a}_i\right\|_2^2\left\|\boldsymbol{b}_j\right\|_2^2\phi_{\max}^2\leq T^2K\left\|\boldsymbol{C}_{S_\delta}\right\|_2^2\left\|\boldsymbol{C}_{S_\delta}(1)\right\|_2^2\phi_{\max}^2,
\end{aligned}\tag{4.A.47}$$

where we use that, by the column normalization on $\boldsymbol{B}_{S_\delta,\perp}$, we have

$$\left\|\boldsymbol{a}_i\right\|_2^2=\left\|\boldsymbol{C}_{S_\delta}'\boldsymbol{\beta}_{S_\delta,\perp,i}\right\|_2^2\leq\left\|\boldsymbol{C}_{S_\delta}\right\|_2^2\left\|\boldsymbol{\beta}_{S_\delta,\perp,i}\right\|_2^2\leq\left\|\boldsymbol{C}_{S_\delta}\right\|_2^2$$

and

$$\|\boldsymbol{b}_j\|_2^2 \leq= \|\boldsymbol{C}_{S_\delta}(1)\boldsymbol{\beta}_{S_\delta,\perp,j}\|_2^2 \leq \|\boldsymbol{C}_{S_\delta}(1)\|_2^2 \|\boldsymbol{\beta}_{S_\delta,\perp,j}\|_2^2 \leq \|\boldsymbol{C}_{S_\delta}(1)\|_2^2 .$$

Plugging (4.A.47) into (4.A.46), we obtain

$$\mathbb{P}\left(\|\boldsymbol{D}_1\|_2 \geq \zeta\right) \leq \frac{s_\delta^4 \|\boldsymbol{C}_{S_\delta}\|_2^2 \|\boldsymbol{C}_{S_\delta}(1)\|_2^2 \phi_{\max}^2}{T^2\zeta^2} \to 0,$$

based on Assumption 4.3 and the assumption that $\frac{s_\delta}{T^{1/2}} \to 0$.

Next, we bound $\boldsymbol{D}_2$ in (4.A.45). By a combination of the union bound, Markov's inequality and Minkowski's inequality,

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{D}_2\|_2 > \zeta\right) &\leq \mathbb{P}\left(\sum_{i,j=1}^{s_\delta}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\boldsymbol{a}_i'\boldsymbol{\epsilon}_{t-1}\boldsymbol{\epsilon}_{t-1}'\boldsymbol{b}_j\right)^2 > \zeta^2\right) \\
&\leq \frac{s_\delta^2 \sum_{i,j=1}^{s_\delta}\mathbb{E}\left(\sum_{t=1}^{T}\sum_{s_1,s_2=1}^{N} a_{i,s_1}b_{j,s_2}\epsilon_{s_1,t-1}\epsilon_{s_2,t-1}\right)^2}{T^4\zeta^2} \\
&\leq \frac{s_\delta^2 \sum_{i,j=1}^{s_\delta}\left(\sum_{t=1}^{T}\sum_{s_1,s_2=1}^{N}|a_{i,s_1}||b_{j,s_2}|\left(\mathbb{E}\left(\epsilon_{s_1,t-1}\epsilon_{s_2,t-1}\right)^2\right)^{1/2}\right)^2}{T^4\zeta^2} \\
&\leq \frac{s_\delta^2 K \sum_{i,j=1}^{s_\delta}\|a_{i,s_1}\|_1^2\|b_{j,s_2}\|_1^2}{T^2\zeta^2} \leq \frac{s_\delta^4 K \|\boldsymbol{C}_{S_\delta}\|_\infty^2\|\boldsymbol{C}_{S_\delta}(1)\|_\infty^2}{T^2\zeta^2} \to 0,
\end{aligned}$$

where we have used that $\mathbb{E}\left(\epsilon_{s_1,t-1}\epsilon_{s_2,t-1}\right)^2 \leq K$ by Assumption 4.1 and the boundedness of $\|\boldsymbol{C}_{S_\delta}\|_\infty^2\|\boldsymbol{C}_{S_\delta}(1)\|_\infty^2$ follows from Assumption 4.3. We omit repeating this argument in the following bounds.

Next, we bound $\boldsymbol{D}_3$. Recall that we define $\boldsymbol{\eta}_t = \boldsymbol{C}_{S_\delta}^*(L)\boldsymbol{\epsilon}_t$, where $\boldsymbol{C}_{S_\delta}^*(z) = \sum_{l=0}^{\infty}\boldsymbol{C}_{S_\delta,l}^* z^l$ with $\sum_{l=0}^{\infty}\left\|\boldsymbol{C}_{S_\delta,l}^*\right\|_\infty \leq K$ by Assumption 4.3. Defining $\boldsymbol{b}_{j,l} = \boldsymbol{C}_{S_\delta,l}^*\boldsymbol{\beta}_{S_\delta,\perp,j}$, we follow a similar strategy to obtain

$$\begin{aligned}
&\mathbb{P}\left(\|\boldsymbol{D}_3\|_2 > \zeta\right) \\
&\leq \frac{s_\delta^2 \sum_{i,j=1}^{s_\delta}\left(\sum_{k=1}^{T-1}\sum_{l=0}^{\infty}\sum_{s_1,s_2=1}^{N}|a_{i,s_1}||b_{j,l,s_2}|\left(\mathbb{E}\left(\epsilon_{s_1,k}\epsilon_{s_2,T-1-l}\right)^2\right)^{1/2}\right)^2}{T^4\phi^2} \\
&\leq \frac{s_\delta^2 K \sum_{i,j=1}^{s_\delta}\left(\sum_{l=0}\|\boldsymbol{b}_{j,l}\|_1\right)^2\|\boldsymbol{a}_i\|_1^2}{T^2\zeta^2} \leq \frac{s_\delta^4 K\left(\sum_{l=0}^{\infty}\left\|\boldsymbol{C}_{S_\delta,l}^*\right\|_\infty\right)^2\|\boldsymbol{C}_{S_\delta}\|_\infty^2}{T^2\zeta^2} \to 0.
\end{aligned}$$

Finally, for $\boldsymbol{D}_4$, we proceed fully analogously, to obtain

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{D}_4\|_2 > \zeta\right) &\leq \frac{s_\delta^2 \sum_{i,j=1}^{s_\delta} \mathbb{E}\left(\sum_{t=1}^{T-1}\sum_{l=0}^{\infty}\sum_{s_1,s_2=1}^{N} a_{i,s_1} b_{j,s_2}\epsilon_{s_1,t}\epsilon_{s_2,t-1-l}\right)^2}{T^4\zeta^2} \\
&\leq \frac{s_\delta^2 \sum_{i,j=1}^{s_\delta}\left(\sum_{t=1}^{T-1}\sum_{l=0}^{\infty}\sum_{s_1,s_2=1}^{N} |a_{i,s_1}|\,|b_{j,s_2}|\left(\mathbb{E}\left(\epsilon_{s_1,t}\epsilon_{s_2,t-1-l}\right)^2\right)^{1/2}\right)^2}{T^4\zeta^2} \\
&\leq \frac{s_\delta^2 K \sum_{i,j=1}^{s_\delta}\left(\sum_{l=0}^{\infty}\|\boldsymbol{b}_j,l\|_1\right)^2\|\boldsymbol{a}_i\|_1^2}{T^2\zeta^2} \leq \frac{s_\delta^4 K\left(\sum_{l=0}^{\infty}\left\|\boldsymbol{C}^*_{S_\delta,l}\right\|_\infty\right)^2\|\boldsymbol{C}_{S_\delta}\|_\infty^2}{T^2\zeta^2} \to 0.
\end{aligned}$$

Combining the results for $\boldsymbol{D}_1$ to $\boldsymbol{D}_4$, it follows that $\mathbb{P}\left(\|\boldsymbol{A}_2\|_2 \geq \zeta\right) \to 0$ as $s_\delta, T \to \infty$.

The last term to derive the stochastic order for is $\boldsymbol{A}_3$ in (4.A.35). Define $\boldsymbol{a}_{i,l} = \boldsymbol{C}'_{S_\delta,l}\boldsymbol{\beta}_{S_\delta,\perp,i}$. Then, again by a combination of the union bound, Markov's inequality and Minkowski's inequality,

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{A}_3\|_2 > \zeta\right) &\leq \mathbb{P}\left(\sum_{i,j=1}^{s_\delta}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\sum_{l_1,l_2=0}^{\infty}\boldsymbol{a}'_{i,l_1}\boldsymbol{\epsilon}_{t-1-l_1}\boldsymbol{\epsilon}'_{t-1-l_2}\boldsymbol{a}_{j,l_2}\right)^2 > \zeta^2\right) \\
&\leq \frac{s_\delta^2 \sum_{i,j=1}^{s_\delta}\mathbb{E}\left(\sum_{t=1}^{T}\sum_{l_1,l_2=0}^{\infty}\sum_{s_1,s_2=1}^{N} a_{i,l_1,s_1}a_{j,l_2,s_2}\epsilon_{s_1,t-1-l_1}\epsilon_{s_2,t-1-l_2}\right)^2}{T^4\zeta^2} \\
&\leq \frac{s_\delta^2 \sum_{i,j=1}^{s_\delta}\left(\sum_{t=1}^{T}\sum_{l_1,l_2=0}^{\infty}\sum_{s_1,s_2=1}^{N}|a_{i,l_1,s_1}|\,|a_{j,l_2,s_2}|\left(\mathbb{E}\left(\epsilon_{s_1,t-1-l_1}\epsilon_{s_2,t-1-l_2}\right)^2\right)^{1/2}\right)^2}{T^4\zeta^2} \\
&\leq \frac{s_\delta^2 K \sum_{i,j=1}^{s_\delta}\left(\sum_{l_1=0}^{\infty}\|a_{i,l_1}\|_1\right)^2\left(\sum_{l_2=0}^{\infty}\|a_{j,l_2}\|_1\right)^2}{T^2\zeta^2} \leq \frac{s_\delta^4 K\left(\sum_{l=0}^{\infty}\|\boldsymbol{C}_{S_\delta,l_1}\|_\infty\right)^4}{T^2\zeta^2}.
\end{aligned}$$

Hence, $\mathbb{P}\left(\|\boldsymbol{A}_3\|_2 \geq \zeta\right) \to 0$ as $s_\delta, T \to \infty$ with $\frac{s_\delta}{T^{1/2}} \to 0$, thereby completing the proof. ■

It is possible to extend the result in Lemma 4.6 to general distributions, based on an argument that relies on strong Gaussian approximations. However, an additional cost is paid in terms of a further restriction on the maximum growth rate of s_δ.

Lemma 4.7. *Let $\hat{\boldsymbol{\Sigma}}_{22}$ be as defined in Assumption 4.4 and maintain Assumptions 4.1-4.3. In addition assume that $\boldsymbol{\epsilon}_t = \boldsymbol{D}\boldsymbol{u}_t$, where $\boldsymbol{D}$ is a T-dimensional square matrix with $\|\boldsymbol{D}\| \leq K$, for some $K > 0$, and $u_{i,s} \perp\!\!\!\perp u_{j,t}$ for all i, j, s, t with $i \neq j$.*

Let $\boldsymbol{\Sigma}_u = (\sigma_{u,ij})_{i,j=1}^N$ *and assume that*

$$\max_{1\leq i\leq N} \mathbb{E}\left|\sum_{t=1}^{T}\left(u_{i,t}^2 - \sigma_{u,ii}^2\right)\right|^2 = O\left(T^{1/2}\right).$$

Then, there exists a constant $\zeta > 0$*, independent of* s_δ*,* N *and* T*, such that*

$$\mathbb{P}\left(\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{22}\right) > \zeta\right) \to 1,$$

as $s_\delta, N, T \to \infty$ *with* $\frac{s_\delta N}{T^{1/4}} \to 0$*.*

Proof. Define $\boldsymbol{s}_{u,t} = \sum_{s=1}^t \boldsymbol{u}_s$, such that $\boldsymbol{s}_t = \boldsymbol{D}\boldsymbol{s}_{u,t}$. The proof makes use of a Gaussian approximation of $\boldsymbol{u}_t$, similar to Zhang et al. (2019a) in their proof of Remark 3.4. By the martingale version of the Skorokhod representation theorem (Strassen, 1967, Theorem 4.3), it is possible to extend the probability space such that, for all i, there exists a standard Brownian motion $W(t)$ and non-negative stopping times $\{\tau_{i,j}\}$ such that for $t \geq 1$,

$$s_{u,it} = W\left(\sum_{j=1}^{t}\tau_{i,j}\right) \text{ and } \mathbb{E}\left[\tau_{i,t}|\mathscr{F}_{i,t-1}\right] = \mathbb{E}\left[u_{i,t}^2|\mathscr{F}_{i,t-1}\right], \tag{4.A.48}$$

where $\mathscr{F}_{i,t}$ is the natural filtration of the stochastic process $\{u_{i,s}, s \leq t\}$. Then, by the proof of Remark 3.4 in Zhang et al. (2019a) it follows that under Assumption 4.1, for every sequence $\{u_{i,t}\}$, there exists an independent and standard normal sequence $\{v_{i,t}\}$, such that

$$\max_{1\leq i\leq N}\max_{0\leq t\leq T}\mathbb{E}\left(\sum_{s=1}^{[Tt]}(u_{i,s} - \sigma_{u,ii}v_{i,s})\right)^2 = O\left(T^{1/2}\right).$$

Define $\tilde{\boldsymbol{s}}_t = \sum_{s=1}^t \tilde{\boldsymbol{\epsilon}}_s$, where $\tilde{\boldsymbol{\epsilon}}_s = (\tilde{\epsilon}_{1,s}, \ldots, \tilde{\epsilon}_{N,s})'$ with $\tilde{\epsilon}_{i,j} = \sigma_{ii}v_{i,j}$. In addition, let

$$\tilde{\boldsymbol{A}}_1 = \boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\boldsymbol{D}\left(\frac{s_\delta}{T^2}\sum_{t=1}^{T}\tilde{\boldsymbol{s}}_t\tilde{\boldsymbol{s}}_t'\right)\boldsymbol{D}'\boldsymbol{C}_{S_\delta}'\boldsymbol{B}_{S_\delta,\perp}.$$

By the proof of Lemma 4.6, there exist a $\zeta > 0$ such that

$$\mathbb{P}\left(\lambda_{\min}\left(\tilde{\boldsymbol{A}}_1\right) > \zeta\right) \to 1,$$

as $s_\delta, T \to \infty$. Recall the decomposition

$$\hat{\boldsymbol{\Sigma}}_{22} = \boldsymbol{A}_1 + \boldsymbol{A}_2 + \boldsymbol{A}_2' + \boldsymbol{A}_3,$$

given by (4.A.35), such that

$$\begin{aligned}\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{22}\right) &\geq \lambda_{\min}\left(\tilde{\boldsymbol{A}}_1\right) + \lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{22} - \tilde{\boldsymbol{A}}_1\right) \\ &\geq \lambda_{\min}\left(\tilde{\boldsymbol{A}}_1\right) - \left\|\boldsymbol{A}_1 - \tilde{\boldsymbol{A}}_1\right\|_2 - 2\left\|\boldsymbol{A}_2\right\|_2 - \left\|\boldsymbol{A}_3\right\|_2.\end{aligned} \tag{4.A.49}$$

In Lemma 4.6, it is shown that for any $\zeta > 0$,

$$\mathbb{P}\left(\left\|\boldsymbol{A}_2\right\|_2 > \zeta\right) \to 0 \text{ and } \mathbb{P}\left(\left\|\boldsymbol{A}_3\right\|_2 > \zeta\right),$$

as $s_\delta, T \to \infty$. Therefore, proving Lemma 4.7 is equivalent to showing that

$$\mathbb{P}\left(\left\|\boldsymbol{A}_1 - \tilde{\boldsymbol{A}}_1\right\|_2 > \zeta\right) \to 0, \tag{4.A.50}$$

as $s_\delta, T \to \infty$ on the extended probability space on which (4.A.48) holds.

We start by deriving the upper bound

$$\begin{aligned}\left\|\boldsymbol{A}_1 - \tilde{\boldsymbol{A}}_1\right\|_2 &\leq= \frac{s_\delta}{T^2}\left\|\boldsymbol{B}_{S_\delta,\perp}'\boldsymbol{C}_{S_\delta}\boldsymbol{D}\left(\sum_{t=1}^{T}\boldsymbol{s}_{u,t}\boldsymbol{s}_{u,t}' - \sum_{t=1}^{T}\tilde{\boldsymbol{s}}_t\tilde{\boldsymbol{s}}_t'\right)\boldsymbol{D}'\boldsymbol{C}_{S_\delta}'\boldsymbol{B}_{S_\delta,\perp}\right\|_2 \\ &\leq \frac{s_\delta}{T^2}\left\|\boldsymbol{D}'\boldsymbol{C}_{S_\delta}'\boldsymbol{B}_{S_\delta,\perp}\right\|_2^2\left\|\sum_{t=1}^{T}\boldsymbol{s}_{u,t}\boldsymbol{s}_{u,t}' - \sum_{t=1}^{T}\tilde{\boldsymbol{s}}_t\tilde{\boldsymbol{s}}_t'\right\|_2,\end{aligned}$$

where

$$\left\|\boldsymbol{D}'\boldsymbol{C}_{S_\delta}'\boldsymbol{B}_{S_\delta,\perp}\right\|_2^2 \leq \left\|\boldsymbol{D}\right\|_2^2\left\|\boldsymbol{C}_{S_\delta}\right\|_2^2\left\|\boldsymbol{B}_{S_\delta,\perp}\right\|_2^2 < \infty,$$

by the assumption that $\|\boldsymbol{D}\|_2 \leq K$, Assumption 4.3 and the normalization on $\boldsymbol{B}_{S_\delta,\perp}$. Furthermore, by the proof of Lemma 9 in the supplementary material of Zhang et al. (2019a, p. 3), it follows that

$$\left\|\sum_{t=1}^{T}\boldsymbol{s}_{u,t}\boldsymbol{s}_{u,t}' - \sum_{t=1}^{T}\tilde{\boldsymbol{s}}_t\tilde{\boldsymbol{s}}_t'\right\|_2 = O_p\left(NT^{7/4}\right),$$

such that

$$\left\|\boldsymbol{A}_1 - \tilde{\boldsymbol{A}}_1\right\|_2 = O_p\left(\frac{s_\delta N}{T^{1/4}}\right).$$

Hence, by (4.A.49) there exist a $\zeta > 0$, such that on an extended probability space

$$\mathbb{P}\left(\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{22}\right) > \zeta\right) \to 1,$$

as $s_\delta, N, T \to \infty$ with $\frac{s_\delta N}{T^{1/4}} \to 0$. ■

Finally, we discuss an alternative route to the result in Lemma 4.6, based on a matrix concentration inequality. Such concentration inequality are becoming increasingly popular in the field of high-dimensional statistics, with excellent recent overviews provided by Tropp (2012, 2015). In particular, we rely on the matrix Chernoff bound bound, the following version of which is stated as Theorem 1.1 in Tropp (2012) and repeated here without proof.

Lemma 4.8. *Consider a finite sequence of $\{\boldsymbol{X}_t\}$ of independent, random, self-adjoint matrices with dimension N. Assume that each random matrix satisfies*

$$\boldsymbol{X}_t \succeq 0, \quad \textit{and} \quad \lambda_{\max}(\boldsymbol{X}_t) \leq R \quad \textit{almost surely.}$$

Define

$$\mu_{\min} := \lambda_{\min}\left(\sum_t \mathbb{E}(\boldsymbol{X}_t)\right).$$

Then, for $\delta \in [0, 1]$,

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{t=1}^{T} \boldsymbol{X}_t\right) \leq (1-\delta)\mu_{\min}\right) \leq N\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu_{\min}/R}. \tag{4.A.51}$$

Based on the Chernoff bound in Lemma 4.8, we derive the following lower bound.

Lemma 4.9. *Define a $(T\times N)$-dimensional matrix with random walks $\boldsymbol{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T)'$, where $\boldsymbol{s}_t = \sum_{s=1}^{t} \boldsymbol{\epsilon}_s$ with $\boldsymbol{\epsilon}_s \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N)$. Assume that $N, T \to \infty$, with $\frac{N}{\log T} \to \infty$ and $\frac{N \log N}{T} \to 0$. Then, there exist a constant $\zeta > 0$, independent of N and T, such that*

$$\mathbb{P}\left(\lambda_{\min}\left(\frac{N \log N}{T^2} \sum_{t=1}^{T} \boldsymbol{s}_t \boldsymbol{s}_t'\right) \leq \zeta\right) \to 0, \tag{4.A.52}$$

as $N, T \to \infty$.

The lower bound in Lemma 4.9 is less tight than the one derived in Lemma 4.6, in the sense that a factor $\log N$ is used to bound the minimum eigenvalue away from

zero. However, Lemma 4.9 requires only $\frac{N \log N}{T} \to 0$ as opposed to the previous $\frac{N}{T^{1/2}} \to 0$. Nonetheless, it turns out that application of Lemma 4.9 to $\lambda_{\min}\left(\hat{\boldsymbol{\Sigma}}_{22}\right)$ does not lead to any improvement over Lemma 4.6, at least not without the use of thus far unknown additional arguments. Therefore, Lemma 4.9 is stated here as a result that may be of independent interest, while not being used for any of the main theorems throughout this chapter.

Proof of Lemma 4.9. As in the proof of Lemma 4.6, we can rewrite $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{E}$, where $\boldsymbol{U}$ is a lower triangular matrix with ones on and below the diagonal and $\boldsymbol{E} = (\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_T)'$. Furthermore, we again decompose $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'$, where $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_T)$ and recall that

$$\lambda_t^{-1} = 4\sin^2\left(\frac{\omega_t}{2}\right) = 2(1 - \cos\omega_t),$$

with $\omega_t = \frac{(2t-1)\pi}{2T+1}$. It follows that

$$\boldsymbol{S}'\boldsymbol{S} = \boldsymbol{E}'\boldsymbol{U}'\boldsymbol{U}\boldsymbol{E} = \tilde{\boldsymbol{E}}'\boldsymbol{\Lambda}\tilde{\boldsymbol{E}} = \sum_{t=1}^{T} \lambda_t \tilde{\boldsymbol{\epsilon}}_t \tilde{\boldsymbol{\epsilon}}_t',$$

where $\tilde{\boldsymbol{\epsilon}}_t \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N)$ by the rotational invariance of the multivariate normal distribution. Consider the matrix $\boldsymbol{S}_\phi'\boldsymbol{S}_\phi = \sum_{t=1}^{T} \lambda_{\phi,t} \tilde{\boldsymbol{\epsilon}}_t \tilde{\boldsymbol{\epsilon}}_t'$, with $\lambda_{\phi,t}^{-1} = 2(1 + \phi_T - \cos\omega_t)$, where ϕ_T is a deterministic function decreasing in T. Note that for any $\phi_T > 0$,

$$\boldsymbol{S}'\boldsymbol{S} - \boldsymbol{S}_\phi'\boldsymbol{S}_\phi = \sum_{t=1}^{T} (\lambda_t - \lambda_{\phi,t}) \tilde{\boldsymbol{\epsilon}}_t \tilde{\boldsymbol{\epsilon}}_t' \succeq 0 \;\Rightarrow\; \lambda_{\min}\left(\boldsymbol{S}'\boldsymbol{S}\right) \geq \lambda_{\min}\left(\boldsymbol{S}_\phi'\boldsymbol{S}_\phi\right). \quad (4.A.53)$$

Let

$$\mu_\phi := \lambda_{\min}\left(\sum_{t=1}^{T} \lambda_{\phi,t}\, \mathbb{E}\left(\tilde{\boldsymbol{\epsilon}}_t \tilde{\boldsymbol{\epsilon}}_t'\right)\right) = \sum_{t=1}^{T} \lambda_{\phi,t},$$

and define $R_\phi = N\lambda_{\phi,1}(1 + \sigma^2)$. We first show that

$$\mathbb{P}\left(\sup_{1 \leq t \leq T} \lambda_{\max}\left(\lambda_{\phi,t} \tilde{\boldsymbol{\epsilon}}_t \tilde{\boldsymbol{\epsilon}}_t'\right) \geq R_\phi\right) \to 0, \quad (4.A.54)$$

as $T, N \to \infty$. First, note that by the union bound, the fact that the $\tilde{\boldsymbol{\epsilon}}_t$ are identically

distributed, the definition of R_ϕ and the triangle inequality, it holds that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{1\leq t\leq T}\lambda_{\max}\left(\lambda_{\phi,t}\tilde{\boldsymbol{\epsilon}}_t\tilde{\boldsymbol{\epsilon}}_t'\right)\geq R_\phi\right)\leq\sum_{t=1}^{T}\mathbb{P}\left(\lambda_{\max}\left(\tilde{\boldsymbol{\epsilon}}_t\tilde{\boldsymbol{\epsilon}}_t'\right)\geq\frac{R_\phi}{\lambda_{\phi,t}}\right)\\
&\leq T\,\mathbb{P}\left(\lambda_{\max}\left(\tilde{\boldsymbol{\epsilon}}_1\tilde{\boldsymbol{\epsilon}}_1'\right)\geq\frac{R_\phi}{\lambda_{\phi,1}}\right)\leq T\,\mathbb{P}\left(\left\|\tilde{\boldsymbol{\epsilon}}_1\right\|_2^2\geq\frac{R_\phi}{\lambda_{\phi,1}}\right)\\
&=T\,\mathbb{P}\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{\epsilon}_{j,1}^2\geq 1+\sigma^2\right)\\
&\leq T\,\mathbb{P}\left(\left|\frac{1}{N}\sum_{j=1}^{N}\tilde{\epsilon}_{j,1}^2-\sigma^2\right|+\sigma^2\geq 1+\sigma^2\right)\\
&=T\,\mathbb{P}\left(\left|\frac{1}{N}\sum_{j=1}^{N}\tilde{\epsilon}_{j,1}^2-\sigma^2\right|\geq 1\right).
\end{aligned}\tag{4.A.55}$$

Since $\tilde{\epsilon}_{j,1}\sim\mathcal{N}(0,1)$ it follows that $\epsilon_{j,1}^2\sim\chi(1)$, i.e. a Chi-squared distribution with one degree of freedom. Moreover, the moment generating function of $\tilde{\epsilon}_{j,1}^2$ is given by (e.g. Casella and Berger, 2002, p. 623),

$$M_\epsilon(t)=\left(\frac{1}{1-2t}\right)^{1/2},\quad\text{for } t<1/2.$$

Since there exists a point K at which $M_\epsilon(K)<\infty$ (e.g. $M_\epsilon(3/8)=2$), it follows from Proposition 2.7.1 (d) in Vershynin (2018) that $\tilde{\epsilon}_{j,1}^2$ is a sub-exponential random variable. Define the sub-exponential norm of a random variable X as

$$\|X\|_{\psi_1}=\inf\left\{t>0:\mathbb{E}\exp\left(|X|/t\right)\leq 2\right\}.\tag{4.A.56}$$

It is straightforward to show that $\tilde{\epsilon}_{j,1}^2-\sigma^2$ is a sub-exponential random variable as well. For the sake of completeness, note that the by definition of $\|\cdot\|_{\psi_1}$, we have $\left\|\sigma^2\right\|_{\psi_1}=\sigma^2\log 2$ and $\left\|\tilde{\epsilon}_{j,1}^2\right\|_{\psi_1}\leq 8/3$, where the latter holds based on the previously stated fact that $M_\epsilon(3/8)=2$. Then,

$$\left\|\tilde{\epsilon}_{j,1}^2-\sigma^2\right\|_{\psi_1}\leq\left\|\tilde{\epsilon}_{j,1}^2\right\|_{\psi_1}+\left\|\sigma^2\right\|_{\psi_1}\leq\left\|\tilde{\epsilon}_{j,1}^2\right\|_{\psi_1}+\sigma^2\log 2\leq 8/3+\sigma^2\log 2=:K_\psi,$$

thereby proving that $\tilde{\epsilon}_{j,1}^2-\sigma^2$ is a sub-exponential random variable. Accordingly, we proceed to bound the RHS of (4.A.55) with the use of a variant of the Bernstein

inequality stated in Corollary 2.8.3. in Vershynin (2018):

$$T\,\mathbb{P}\left(\left|\frac{1}{N}\sum_{j=1}^{N}\tilde{\epsilon}_{j,1}^2-\sigma^2\right|\geq 1\right)\leq 2T\exp\left(-\frac{cN}{K_\psi^2}\right)\to 0. \tag{4.A.57}$$

as $T, N \to \infty$ and under the assumption that $\frac{N}{\log T} \to \infty$. This confirms the claim in (4.A.54).

Continuing with deriving the minimum eigenvalue bound, we let

$$K_\delta = \frac{e^{-\delta}}{(1-\delta)^{1-\delta}},$$

such that $0 < K_\delta < 1$ for any $\delta \in (0,1)$. By the matrix Chernoff bound in Lemma 4.8, it holds that

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{t=1}^{T}\lambda_{\phi,t}\tilde{\boldsymbol{\epsilon}}_t\tilde{\boldsymbol{\epsilon}}_t'\right)\leq(1-\delta)\mu_\phi\right)\leq NK_\delta^{\mu_\phi/R_\phi}. \tag{4.A.58}$$

Hence, our goal is to ensure that μ_ϕ diverges as fast as possible via our choice of ϕ_T, while simultaneously ensuring that $\mu_\phi/R_\phi \to \infty$. First, we derive a lower bound for μ_ϕ as a function of ϕ_T as

$$\begin{aligned}2\mu_\phi &= \sum_{t=1}^{T}\frac{1}{1-\cos\omega_t+\phi_T}\geq\sum_{t=1}^{T}\frac{1}{\left(\frac{(2t+1)\pi}{2T+1}\right)^2+\phi_T}\\ &= \sum_{t=1}^{T}\frac{(2T+1)^2}{(2t+1)^2\pi^2+\phi_T(2T+1)^2}.\end{aligned} \tag{4.A.59}$$

Let a_T be a slowly increasing function such that $a_T \to \infty$ with $\frac{Na_T}{T} \to 0$. Note that by the assumption that $\frac{N\log N}{T} \to 0$, we may set $a_T = \log N^K$ for any $K > 0$. Next, set $\phi_T^{1/2} = \frac{Na_T}{T}$ and define $[x]$ as the integer part of x. Then, for large enough T we

have $\phi_T^{1/2} \leq 1$, such that

$$\begin{aligned}
\mu_\phi &\geq \sum_{t=1}^{T} \frac{(2T+1)^2}{2\left((2t+1)^2\pi^2 + \phi_T(2T+1)^2\right)} \\
&\geq \sum_{t=1}^{T} \frac{(2T)^2}{2\left((3t)^2\pi^2 + \phi_T(3T)^2\pi^2\right)} \\
&= \sum_{t=1}^{T} \frac{4T^2}{18\pi^2\left(t^2 + \phi_T T^2\right)} \geq \sum_{t=1}^{\left[T\phi_T^{1/2}\right]} \frac{4T^2}{18\pi^2\left(t^2 + \phi_T T^2\right)} \\
&\geq \frac{4\phi_T^{1/2}T^3}{36\pi^2\phi_T T^2} = \frac{T}{9\pi^2\phi_T^{1/2}} = \frac{T^2}{9\pi^2 N a_T} \to \infty.
\end{aligned} \tag{4.A.60}$$

Next, we focus on μ_ϕ / R_ϕ. Recall the definition $R_\phi = (1+\sigma^2)\lambda_{\phi,1}N$. Then,

$$\frac{\mu_\phi}{R_\phi} \geq \frac{T}{9\pi^2(1+\sigma^2)\lambda_{\phi,1}N\phi_T^{1/2}} \geq \frac{T\phi_T^{1/2}}{9\pi^2(1+\sigma^2)N} = \frac{a_T}{9\pi^2} \to \infty. \tag{4.A.61}$$

Finally, setting $K_\delta = \frac{e^{-\delta}}{(1-\delta)^{1-\delta}}$, for some $\delta \in (0,1)$, and recalling that $a_T = K_a \log N$, where we now define $K_a > \frac{9\pi^2}{\log K_\delta}$. Then, combining (4.A.53), (4.A.54), (4.A.60) and (4.A.61), and using the Chernoff matrix bound in Lemma 4.8, we conclude that

$$\begin{aligned}
&\mathbb{P}\left(\lambda_{\min}\left(\frac{N\log N}{T^2}\boldsymbol{S}'\boldsymbol{S}\right) \leq \frac{1-\delta}{K_a 9\pi^2}\right) = \mathbb{P}\left(\lambda_{\min}\left(\frac{9\pi^2 N a_T}{T^2}\boldsymbol{S}'\boldsymbol{S}\right) \leq 1-\delta\right) \\
&\leq \mathbb{P}\left(\lambda_{\min}\left(\frac{1}{\mu_\phi}\boldsymbol{S}_\phi'\boldsymbol{S}_\phi\right) \leq 1-\delta\right) \\
&\leq \mathbb{P}\left(\lambda_{\min}\left(\frac{1}{\mu_\phi}\boldsymbol{S}_\phi'\boldsymbol{S}_\phi\right) \leq 1-\delta, \sup_{1\leq t\leq T}\lambda_{\max}\left(\lambda_{\phi,t}\tilde{\boldsymbol{\epsilon}}_t\tilde{\boldsymbol{\epsilon}}_t'\right) \leq R_\phi\right) \\
&\quad + \mathbb{P}\left(\sup_{1\leq t\leq T}\lambda_{\max}\left(\lambda_{\phi,t}\tilde{\boldsymbol{\epsilon}}_t\tilde{\boldsymbol{\epsilon}}_t'\right) > R_\phi\right) \leq N K_\delta^{\mu_\phi/R_\phi} + o(1) \leq N K_\delta^{a_T/9\pi^2} + o(1) \\
&= N\exp\left(\frac{a_T\log K_\delta}{9\pi^2}\right) + o(1) \to 0,
\end{aligned}$$

by our assumption on K_a. Hence, we have proven Lemma 4.9 for $\zeta = \frac{1-\delta}{K_a 9\pi^2}$. ■

Chapter 5

High-dimensional Forecasting in the Presence of Unit Roots and Cointegration

"The modern macro economy is large, diffuse, and difficult to define, measure, and control."

- C. Granger (1934-2009)

Abstract[†]

In this chapter we investigate how the possible presence of unit roots and cointegration affects forecasting based on high-dimensional datasets. When modelling (co)integrated data, the researcher is required to either transform the integrated time series to stationarity or to explicitly model the cointegrating properties of the data. However, both approaches are complicated by the high-dimensional setting. First, transformations to stationarity require performing many unit root tests, increasing room for errors in the classification. Second, modelling unit roots and cointegration directly is more difficult, as standard high-dimensional techniques such as factor models and penalized regression are not directly applicable to (co)integrated data and need to be adapted. In this chapter we provide an overview of both issues and review methods proposed to address these issues. These methods are also illustrated with two empirical applications.

[†]This chapter is based on Smeekes and Wijler (2020).

5.1 Introduction

In this chapter we investigate forecasting based on high-dimensional datasets in which the series may contain unit roots and be cointegrated. As most macroeoconomic time series are at least very persistent, and may contain unit roots, a proper handling of unit roots and cointegration is of paramount importance in macroeconomic forecasting. The theory of unit roots and cointegration in small systems is well-developed and numerous reference works exist to guide the practitioner, see for example Enders (2008) or Hamilton (1994) for comprehensive treatments.

We discuss the problems that arise when extending the analysis to high-dimensional data and consider solutions that have been proposed in the literature. In particular, we discuss the applicability of the proposed methods for macroeconomic forecasting, reviewing relevant theoretical properties and practical issues. Moreover, by reconsidering the two high-dimensional applications of Chapters 2 and 3 —which are very different in spirit— we illustrate the issues and analyze the performance of the various methods in practically relevant situations.

The empirical literature dealing with unit roots and cointegration can essentially be split into two different philosophies. The first approach is to apply an appropriate transformation to each series such that one can work with stationary time series, with the most common transformation taking first differences of a series with a unit root. This is the most common approach in high-dimensional forecasting, as it only involves 'straightforward' unit root or stationarity testing on each series. Indeed, commonly used high-dimensional datasets such as the FRED-MD and -QD datasets (McCracken and Ng, 2016) already come with pre-determined transformation codes to achieve stationarity.[1] While this approach appears to be conceptually simple, we will argue in this chapter that there are seemingly minor issues that are often ignored in practice, but which can have a big impact on the performance of consequent forecasts, in particular when working with less established datasets.

The second approach is to model unit root and cointegration properties directly. In small systems, this is commonly done through vector error correction models (VECM), often using the popular maximum likelihood methodology developed by Johansen (1995a). The rationale for this seemingly more complicated approach is that ignoring long-run relations between the variables, as is done in the first approach, means not incorporating all information into the forecaster's model, which may have a detrimental

[1] The transformations to stationarity in the empirical application of Chapter 2 were based on these pre-determined transformation codes.

effect on the forecast quality. Extending these techniques for modelling cointegration to high-dimensional settings requires a careful rethink of how cointegration can be viewed in high dimensions, and is an ongoing area of research. We will discuss recent contributions in this area and analyze the respective merits and drawbacks of each method.

While the importance of the concept of cointegration for macroeconometric analysis cannot be understated, one might argue that for the specific goal of forecasting it is not crucial. In the low-dimensional time series literature a large body of literature exists which compares the relative merits of the two philosophical approaches for forecasting, see for instance Clements and Hendry (1995), Christoffersen and Diebold (1998), Diebold and Kilian (2000) and the references therein. Generally, the conclusion is mixed, with the performance of each approach varying depending on forecast horizon, dimensions of the models, estimation accuracy, and even specific applications and datasets. As this is no different in a high-dimensional context, we make no attempt to classify one of these approaches as superior. Instead, we aim to provide the practitioner with an overview of tools available to follow either line of thought.

One could discern a third approach to unit roots and cointegration, which is to ignore unit roots all together and estimate all forecasting models in levels. While this approach is at first glance close to the first approach and one might have valid reasons to prefer this approach, we do not recommend this in high-dimensional problems. If cointegration is not present in (parts of) the data, these methods may be very sensitive to spurious regression. The higher the dimensions of the data, the more likely that spurious regression becomes an issue. In particular, given that many methods discussed in this book perform some sort of dimensionality reduction or variable selection, this may actually increase the likelihood of obtaining spurious results. For instance, we observed in Chapter 2 that the variable selection of lasso-type estimators quickly deteriorated when naively applied to a mix of cointegrated and spurious regressors. Low-dimensional solutions such as always including lagged levels to avoid spurious regression are not possible in high-dimensional systems, as it would require including too many variables, and the applied dimensionality reduction or variable selection techniques might not be able to retain the lagged levels in the model. As such, we do not consider the approach of estimating everything in levels further in this chapter.[2]

We also illustrate the discussed methods by two empirical applications. In the

[2]Obviously, this caveat does not mean that forecasting in levels does not yield good results for specific applications. The applied researcher is free to apply any of the methods discussed in this chapter directly to (suspected) unit root series, but should simply be wary of the results.

first we forecast several U.S. macroeconomic variables using the FRED-MD database, similar to the application considered in Chapter 2. This application tests the methods in a known macroeconomic context, thus serving as a benchmark. In our second application, we consider nowcasting unemployment using a dataset constructed from Google Trends with frequencies of unemployment-related search terms, similar to the application considered in Chapter 3. This second application not only serves to highlight the potential of 'modern' sources for high-dimensional datasets by which to forecasting macroeconomic time series, but also illustrates that in such applications, we have little theoretical guidance to decide on unit root and cointegration properties, and proper data-driven methods are needed.

Note that, as is common in the related high-dimensional literature, we focus explicitly on point forecasts. As distributional theory changes when unit roots are present, performing interval forecasts in the presence of unit roots and cointegration is a much more challenging – and largely unresolved – issue in the high-dimensional setting, especially as it adds to the complications of performing inference in high dimensions already present without unit roots. Given the sparsity of literature on this topic, we do not consider interval prediction in this chapter. This is clearly a very important avenue for future research.

The remainder of this chapter is organized as follows. Section 5.2 describes the general setup and introduces the cointegration model, along with some useful representations for later use. We discuss how to transform high-dimensional datasets to stationarity in Section 5.3, while Section 5.4 introduces high-dimensional approaches for modelling cointegration. In Section 5.5 we apply the discussed methods to our two empirical forecasting exercises. Finally, Section 5.6 concludes.

5.2 General Setup

In this section we describe a general model for cointegration to be used throughout the chapter. Next to defining the model in the classical error correction form, we also consider alternative representations that will be useful later in the chapter.

Let $\boldsymbol{z}_t$ denote an N-dimensional time series observed at time $t = 1, \dots, T$. Assume that we can represent the series as

$$\boldsymbol{z}_t = \boldsymbol{\mu} + \boldsymbol{\tau} t + \boldsymbol{\zeta}_t, \tag{5.2.1}$$

where $\boldsymbol{\mu}$ is an N-dimensional vector of intercepts, $\boldsymbol{\tau}$ is an n-dimensional vector of

trend slopes, and $\boldsymbol{\zeta}_t$ is the N-dimensional purely stochastic time series. This stochastic component given is by

$$\Delta\boldsymbol{\zeta}_t = \boldsymbol{A}\boldsymbol{B}'\boldsymbol{\zeta}_{t-1} + \sum_{j=1}^{p} \boldsymbol{\Phi}_j \Delta\boldsymbol{\zeta}_{t-j} + \boldsymbol{\varepsilon}_t, \tag{5.2.2}$$

where $\boldsymbol{\varepsilon}_t$ is the N-dimensional innovation vector. Generally the innovations $\boldsymbol{\varepsilon}_t$ will be a martingale difference sequence, although we abstract from making too specific assumptions at this point.

We can obtain the classical vector error correction model (VECM) for $\boldsymbol{z}_t$ by substituting (5.2.1) into (5.2.2):

$$\Delta\boldsymbol{z}_t = \boldsymbol{A}\boldsymbol{B}' \left(\boldsymbol{z}_{t-1} - \boldsymbol{\mu} - \boldsymbol{\tau}(t-1)\right) + \boldsymbol{\tau}^* + \sum_{j=1}^{p} \boldsymbol{\Phi}_j \Delta\boldsymbol{z}_{t-j} + \boldsymbol{\varepsilon}_t, \tag{5.2.3}$$

where $\boldsymbol{\tau}^* = (\boldsymbol{I}_N - \sum_{j=1}^{p} \boldsymbol{\Phi}_j)\boldsymbol{\tau}$. The long-run relations are contained in the $(N \times r)$-matrix $\boldsymbol{B}$, while the $(N \times r)$ matrix $\boldsymbol{A}$ contains the corresponding loadings. Here the variable r describes the number of cointegrating relations in the systems. If $r = 0$, we adopt the convention that $\boldsymbol{A}\boldsymbol{B}' = 0$; in this case $\boldsymbol{z}_t$ is a pure N-dimensional unit root process. If $r = N$, all series are $I(0)$. To ensure that $\boldsymbol{z}_t$ is at most an $I(1)$ process, the lag polynomial $\boldsymbol{C}(z) := (1-z) - \boldsymbol{A}\boldsymbol{B}'z - \sum_{j=1}^{p} \boldsymbol{\Phi}_j (1-z) z^j$ and matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ should satisfy Assumption 3.2. Under this assumptions, exactly $N - r$ roots of the lag polynomial $\boldsymbol{C}(z)$ are equal to unity, while the remaining r roots lie outside the unit circle.

From the Granger Representation Theorem (cf. Johansen, 1995a, p. 49), we can obtain the *common trend representation* of (5.2.3), which is given by

$$\boldsymbol{z}_t = \boldsymbol{\mu} + \boldsymbol{\tau}t + \boldsymbol{C}\boldsymbol{s}_t + \boldsymbol{u}_t, \tag{5.2.4}$$

where $\boldsymbol{C}$ is an $(N \times N)$ matrix of rank $N - r$,[3] $\boldsymbol{s}_t = \sum_{i=1}^{t} \boldsymbol{\varepsilon}_t$ are the stochastic trends and $\boldsymbol{u}_t$ is a stationary process. This representation show that $\boldsymbol{z}_t$ can be decomposed in a deterministic process, an $I(1)$ part of common trends, $\boldsymbol{C}\boldsymbol{s}_t$, and a stationary part $\boldsymbol{u}_t$.

To see the commonality of the trends, note that as $\boldsymbol{C}$ is of reduced rank, we can define $(N \times (N-r))$ matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$ such that $\boldsymbol{C} = \boldsymbol{\Lambda}\boldsymbol{\Gamma}'$. Then defining the

[3] If $r = 0$, we set $\boldsymbol{C} = \boldsymbol{0}$.

$((N-r)\times 1)$-vector $\boldsymbol{f}_t = \boldsymbol{\Gamma}'\boldsymbol{s}_t$, we can write (5.2.4) as

$$\boldsymbol{z}_t = \boldsymbol{\mu} + \boldsymbol{\tau}t + \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{u}_t. \tag{5.2.5}$$

We can now see the common trends as *common factors*, which provides a convenient way to think about cointegration in high dimensions.

This brings us to an alternative way to represent cointegration through a common factor structure from the outset. This form was considered by Bai and Ng (2004) among others to investigate different sources of nonstationarity in a panel data context. In this case we start from (5.2.5), assuming that the elements of both $\boldsymbol{f}_t$ and $\boldsymbol{u}_t$ can be $I(0)$ or $I(1)$. The combination of the two then determines the properties of the series $\boldsymbol{z}_t$. Consider a single series $\boldsymbol{z}_{i,t}$, which can be represented as

$$z_{i,t} = \mu_i + \tau_i t + \boldsymbol{\lambda}_i'\boldsymbol{f}_t + u_{i,t},$$

where $\boldsymbol{\lambda}_i'$ denotes the i-th row of $\boldsymbol{\Lambda}$. Note that $\boldsymbol{z}_{i,t}$ is $I(0)$ only if both $\boldsymbol{u}_{i,t}$ and $\boldsymbol{\lambda}_i'\boldsymbol{f}_t$ are $I(0)$, where the latter occurs if either all factors $\boldsymbol{f}_t$ are $I(0)$, or no $I(1)$ factors load on series i. Similarly, cointegration between series i and j requires that both $u_{i,t}$ and $u_{j,t}$ are $I(0)$.

Remark 5.1. For expositional simplicity we do not consider $I(2)$ variables here. While the VECM can be extended to allow for $I(2)$ series, see e.g. Johansen (1995b), in practice most cointegration analyses are performed on $I(1)$ series. If the data contains (suspected) $I(2)$ series, these are generally differenced before commencing the cointegration analysis.

Similarly, one could think of the data generating process (DGP) as being of infinite lag order, rather than fixed order p. In this case the VECM with fixed order can be thought of as an approximation to the infinite order model, where p should be large enough to capture 'enough' of the serial correlation. Either way, in applications p is generally not known and has to be estimated.

5.3 Transformations to Stationarity and Unit Root Pre-Testing

In this section we discuss how to determine the appropriate transformations —in particular how often the series need to be differenced— in order to obtain only stationary time series in our dataset. While established datasets, such as the FRED-MD,

come with an overview of the appropriate transformation for each series, this is generally not the case and data-driven methods are needed. Thus, one normally has to apply unit root or stationarity tests to determine the order of integration, and the corresponding transformation. In this section we investigate how to approach this pre-testing problem.

First, we investigate unit root tests in more detail, and highlight some of their characteristics that one should take into account when considering high-dimensional macroeconomic forecasting. Second, we discuss how to deal with the multiple testing problem that arises from the fact that we need to combine unit root tests on many time series.

5.3.1 Unit Root Test Characteristics

Even though the literature on unit root testing has grown exponentially since the seminal paper of Dickey and Fuller (1979), discussing at length the characteristics of various unit root tests, unit root pre-testing is often done in an automatic, routine-like, way by considering classical tests such as augmented Dickey-Fuller (ADF) tests. However, these tests have various problematic characteristics which may accumulate when applied in high-dimensional problems. While we cannot discuss all of these here, let us briefly mention some of particular relevance for macroeconomic forecasting. An extensive overview of unit root testing is provided by Choi (2015).[4]

Size distortions

Standard unit root tests are very prone to size distortions. One source is neglected serial correlation (cf. Schwert, 1989), while another is time-varying volatility (Cavaliere, 2005). For both sources, bootstrap methods have proven a successful means to counteract the size distortions; however, while for serial correlation any 'off-the-shelf' time series bootstrap method can be used (see Palm et al., 2008, for an overview and comparison), dealing with general forms of heteroskedasticity requires a unit root test based on the wild bootstrap (Cavaliere and Taylor, 2008, 2009).

It should be noted that unconditional volatility changes pose a particular concern for macroeconomic time series. Many datasets such as FRED-MD span the period of the Great Moderation, which has significantly affected the volatility of macroeconomic time series (Justiniano and Primiceri, 2008; Stock and Watson, 2003). It would

[4]Given the greater popularity of tests where the null hypothesis is a unit root over tests with stationarity as the null, we focus exclusively on unit root tests here. However, most of the discussion applies to stationarity tests as well.

therefore appear wise to take potential volatility changes into account when selecting an appropriate unit root test.

Power and specification considerations

The power properties of the different unit root tests proposed vary considerably, and generally optimal tests do not exist. One particular source of variation is the magnitude of the initial condition, where for instance the DF-GLS test of Elliott et al. (1996) is optimal when the initial condition is zero, but the ADF test is much more powerful when the initial condition is large (Müller and Elliott, 2003). An even larger source of variation is the presence or absence of a deterministic trend. Unit root tests with a trend included (or, equivalently, unit root tests performed on detrended data) are considerably less powerful than without trend (performed on demeaned data). On the other hand, if a trend is not included when the data do contain one, the unit root test is not correctly sized anymore (Harvey et al., 2009).

While dealing with such issues is manageable in unit root testing for a single series, this changes when considering large datasets. For instance, deciding whether to include a trend in the unit root test can be based on a combination of theory, visual inspection, pre-testing, and comparing outcomes of different tests with or without a trend. However, such an analysis has to be done manually for each series involved, which quickly becomes problematic if the dimension of the dataset increases. This is even more problematic for modern high-dimensional datasets, such as Google Trends, for which no theory exists to guide the practitioner, and where the dimension can become arbitrarily large.

As such one would like to have an automatic way of choosing good specifications for the unit root tests, that may differ across series. One easy way is provided by the union of unit root tests principle proposed by Harvey et al. (2009, 2012), in which several unit root tests are performed, and the unit root null hypothesis is rejected if one of the tests rejects (when corrected for multiple testing). In particular, Harvey et al. (2012) consider a union of the ADF and DF-GLS tests, both with and without linear trend, to cover uncertainty about both trend and initial condition. Smeekes and Taylor (2012) consider a wild bootstrap version of this test that is robust to

time-varying volatility. The test statistic for series i takes the form

$$UR_i = \min\left(\left(\frac{x_i}{c^{\mu*}_{i,GLS}(\alpha)}\right) GLS^{\mu}_i, \left(\frac{x_i}{c^{\tau*}_{i,GLS}(\alpha)}\right) GLS^{\tau}_i, \left(\frac{x_i}{c^{\mu*}_{i,ADF}(\alpha)}\right) ADF^{\mu}_i, \left(\frac{x_i}{c^{\tau*}_{i,ADF}(\alpha)}\right) ADF^{\tau}_i\right), \tag{5.3.1}$$

where ADF_i and GLS_i are the ADF and DF-GLS test performed on series i, while superscript μ and τ indicate whether the series are demeaned or detrended respectively. The bootstrap critical values such as $c^{\mu*}_{i,GLS}(\alpha)$ used in the scaling factors are determined in a preliminary bootstrap step as the individual level α critical values of the four tests. The variable x_i is a scaling factor to which the statistics are scaled. Any $x_i < 0$ suffices to preserve the left-tail rejection region; if one additionally takes x_i the same value for all series i, test statistics become comparable across series, which facilitates the multiple comparisons discussed in the next subsection.

5.3.2 Multiple Unit Root Tests

Performing a unit root test for every series separately raises issues associated with multiple testing. In particular, the probability of incorrect classifications rises with the number of tests performed. If each test has a significance level of 5%, we may also expect roughly 5% of the $I(1)$ series to be incorrectly classified as $I(0)$. In a high-dimensional dataset this can quickly lead to a significant number of incorrectly classified series. It will of course depend on the specific application whether this is problematic —a priori we cannot say whether the 'important' series will be correctly classified or not— but to avoid such issues one can formally account for multiple testing.

There is a huge statistical literature about multiple testing; Romano et al. (2008b) provide an overview with a focus on econometric applications. Here we briefly discuss the most prominent methods developed for the purposes of unit root testing. Before discussing the different methods to control for multiple testing, let us set up the general framework. Let $UR_1, \ldots, UR_N$ denote the unit root test statistics for series 1 up to N, assuming they reject for small values of the statistics.[5] It is important to choose the test statistics such that they are directly comparable, in the sense that their marginal distributions are the same. If this is the case, then the ranking

$$UR_{(1)} \leq \ldots \leq UR_{(R)} \leq UR_{(R+1)} \leq \ldots \leq UR_{(N)}, \tag{5.3.2}$$

[5] We can assume this without loss of generality as any test statistic can be modified to indeed do so.

where $UR_{(i)}$ denotes the i-th order statistic of $UR_1, \ldots, UR_N$, corresponds to a ranking from 'most significant' to 'least significant'. To ensure the comparability of the test statistics, one needs to eliminate nuisance parameters from their distribution. Hence, simply using the bootstrap to absorb nuisance parameters is not sufficient; instead, one often needs to transform (for instance to p-values) or scale the statistics appropriately. In the union tests of (5.3.1), the scaling is done automatically by setting $x_i = -1$ for all units.

Given the ranking in (5.3.2), the objective is to find an appropriate cut-off point R such that for all statistics less than or equal $UR_{(R)}$ the unit root hypothesis is rejected, and for all statistics larger it is not rejected. How this threshold is determined depends on how multiple testing is controlled for.

Controlling generalized error rates

Generalized error rates provide multivariate extensions of the standard Type I error. The most common is the *familywise error rate (FWE)*, which is defined as the probability of making at least one false rejection of the null hypothesis. This can easily be controlled by the popular Bonferroni correction. However, this is very conservative as it is valid under any form of dependence. On the contrary, if the bootstrap is used to capture the actual dependence structure among the tests, one can control for multiple testing without the need for being conservative. This approach is followed by Hanck (2009), who controls FWE in unit root testing by applying the bootstrap algorithm proposed by Romano and Wolf (2005).

While controlling FWE makes sense when N is small, in typical high-dimensional datasets FWE becomes too conservative. Instead, one can control the *false discovery rate (FDR)* originally proposed by Benjamini and Hochberg (1995), which is defined as

$$FDR = \mathbb{E}\left[\frac{F}{R}\mathbb{1}(R > 0)\right],$$

where R denote the total number of rejections, and F the number of false rejections. The advantage of the FDR is that it scales with increasing N, and thus is more appropriate for large datasets. However, most non-bootstrap methods are either not valid under arbitrary dependence or overly conservative. Moon and Perron (2012) compare several methods to control FDR and find that the bootstrap method of Romano et al. (2008a), hereafter denoted as BFDR, does not share these disadvantages and clearly outperforms the other methods. A downside of this method however is

that the algorithm is rather complicated and time-consuming to implement. Globally, the algorithm proceeds in a sequential way by starting to test the 'most significant' series, that is, the smallest unit root test statistic. This statistic is then compared to an appropriate critical values obtained from the bootstrap algorithm, where the bootstrap evaluates all scenarios possible in terms of false and true rejections given the current progression of the algorithm. If the null hypothesis can be rejected for the current series, the algorithm proceeds to the next most significant statistic and the procedure is repeated. Once a non-rejection is observed, the algorithm stops. For details we refer to Romano et al. (2008a). This makes the bootstrap FDR method a *step-down* method, contrary to the original Benjamini and Hochberg (1995) approach which is a step-up method starting from the least significant statistic.

Sequential testing

Smeekes (2015) proposes an alternative bootstrap method for multiple unit root testing based on sequential testing. In a first step, the null hypothesis that all N series are $I(1)$ –hence $p_1 = 0$ series are $I(0)$— is tested against the alternative that (at least) p_2 series are $I(0)$. If the null hypothesis is rejected, the p_2 most significant statistics in (5.3.2) are deemed $I(0)$ and removed from consideration. Then the null hypothesis that all remaining $N - p_2$ series are $I(1)$ is tested against the alternative that at least p_2 of them are $I(0)$, and so on. If no rejections are observed, the final rounds tests p_K $I(0)$ series against the alternative of N $I(0)$ series. The numbers $p_2, \dots, p_K$ as well as the number of tests K are chosen by the practitioner based on the specific application at hand. By choosing the numbers as $p_k = [q_k N]$, where $q_1, \dots, q_K$ are desired quantiles, the method automatically scales with N.

Unlike the BFDR method, this Bootstrap Sequential Quantile Test (BSQT) is straightforward and fast to implement. However, it is dependent on the choice of numbers p_k to be tested; its 'error allowance' is therefore of a different nature to error rates like FDR. Smeekes (2015) shows that, when p_J units are found to be $I(0)$, the probability that the true number of $I(0)$ series lies outside the interval $[p_{J-1}, p_{J+1}]$ is at most the chosen significance level of the test. As such, there is some uncertainty around the cut-off point.

It might therefore be tempting to choose $p_k = k - 1$ for all $k = 1, \dots, N$, such that this uncertainty disappears. However, as discussed in Smeekes (2015), applying the sequential method to each series individually hurts power if N is large as it amounts to controlling FWE. Instead, a better approach is to iterate the BSQT method; that is, it can be applied in a second stage just to the interval $[p_{J-1}, p_{J+1}]$

to reduce the uncertainty. This can be iterated until few enough series remain to be tested individually in a sequential manner. On the other hand, if $p_1, \ldots, p_K$ are chosen sensibly and not spaced too far apart, the uncertainty is limited to a narrow range around the 'marginally significant' unit root tests. These series are at risk of missclassification anyway, and the practical consequences of incorrect classification for these series on the boundary of a unit root are likely small.

Smeekes (2015) performs a Monte Carlo comparison of the BSQT and BFDR methods, as well as several methods proposed in the panel data literature such as Ng (2008) and Chortareas and Kapetanios (2009). Globally BSQT and BFDR clearly outperform the other methods, where BFDR is somewhat more accurate than BSQT when the time dimension T is at least of equal magnitude as the number of series N. On the other hand, when T is much smaller than N BFDR suffers from a lack of power and BSQT is clearly preferable. In our empirical applications we will therefore consider both BFDR and BSQT, as well as the strategy of performing individual tests without controlling for multiple testing.

Remark 5.2. An interesting non-bootstrap alternative is the panel method proposed by Pedroni et al. (2015), which has excellent performance in finite samples. However, implementation of this method requires that T is strictly larger than N, thus severely limiting its potential in the high-dimensional setting. Another alternative would be to apply the model selection approach through the adaptive lasso by Kock (2016) which avoids testing all together. However, this has only been proposed in a univariate context and its properties are unknown for the type of application considered here.

Multivariate bootstrap methods

All multiple testing methods described above require a bootstrap method that can not only account for dependence within a single time series, but can also capture the dependence structures between series. Accurately modelling the dependence between the individual test statistics is crucial for proper functioning of the multiple testing corrections. Capturing the strong and complex dynamic dependencies between macroeconomic series requires flexible bootstrap methods that can handle general forms of dependence.

Moon and Perron (2012) and Smeekes (2015) use the moving-blocks bootstrap (MBB) based on the results of Palm et al. (2011) who prove validity for mixed $I(1)/I(0)$ panel datasets under general forms of dependence. However the MBB has two disadvantages. First, it can only be applied to balanced datasets where each time series is observed over the same period. This makes application to datasets such

as FRED-MD difficult, at least without deleting observations for series that have been observed for a longer period. Second, the MBB is sensitive to unconditional heteroskedasticity, which makes its application problematic for series affected by the Great Moderation.

Dependent wild bootstrap (DWB) methods address both issues while still being able to capture complex dependence structure. Originally proposed by Shao (2010) for univariate time series, they were extended to unit root testing by Smeekes and Urbain (2014b) and Rho and Shao (2019), where the former paper considers the multivariate setup needed here. A general wild bootstrap algorithm for multivariate unit root testing looks as follows:

1. Detrend the series $\{\boldsymbol{z}_t\}$ by OLS; that is, let $\hat{\boldsymbol{\zeta}}_t = (\widehat{\zeta}_{1,t}, \dots, \widehat{\zeta}_N)'$ where

$$\widehat{\zeta}_{i,t} = z_{i,t} - \widehat{\mu}_i - \widehat{\tau}_i t, \qquad i = 1, \dots, N, \quad t = 1, \dots, T$$

   and $(\widehat{\mu}_i, \widehat{\tau}_i)'$ are the OLS estimators of $(\mu_i, \tau_i)'$.

2. Transform $\widehat{\boldsymbol{\zeta}}_t$ to a multivariate $I(0)$ series $\widehat{\boldsymbol{u}}_t = (\widehat{u}_{1,t}, \dots, \widehat{u}_{N,t})'$ by setting

$$\widehat{u}_{i,t} = \widehat{\zeta}_{i,t} - \widehat{\rho}_i \widehat{\zeta}_{i,t-1}, \qquad i = 1, \dots, N, \quad t = 1, \dots, T,$$

   where $\widehat{\rho}_i$ is either an estimator of the largest autoregressive root of $\{\widehat{\zeta}_{i,t}\}$ using for instance an (A)DF regression, or $\widehat{\rho}_i = 1$.

3. Generate a univariate sequence of *dependent* random variables $\xi_1^*, \dots, \xi_N^*$ with the properties that $\mathbb{E}^* \xi_t^* = 0$ and $\mathbb{E}^* \xi_t^{*2} = 1$ for all t. Then construct bootstrap errors $\boldsymbol{u}_t^* = (u_{1,t}^*, \dots, u_{N,t}^*)'$ as

$$u_{i,t}^* = \xi_t^* \hat{u}_{i,t}, \qquad i = 1, \dots, N, \quad t = 1, \dots, T. \tag{5.3.3}$$

4. Let $\boldsymbol{z}_t^* = \sum_{s=1}^t \boldsymbol{u}_s^*$ and calculate the desired unit root test statistics $UR_1^*, \dots, UR_N^*$ from $\{\boldsymbol{z}_t^*\}$. Use these bootstrap test statistics in an appropriate algorithm for controlling multiple testing.

Note that, unlike for the MBB, in (5.3.3) no resampling takes place, and as such missing values 'stay in their place' without creating new 'holes' in the bootstrap samples. This makes the method applicable to unbalanced panels. Moreover, heteroskedasticity is automatically taken into account by virtue of the wild bootstrap principle. Serial dependence is captured through the dependence of $\{\xi_t^*\}$, while dependence across series is captured directly by using the same, univariate, ξ_t^* for each

series i. Smeekes and Urbain (2014b) provide theoretical results on the bootstrap validity under general forms of dependence and heteroskedasticity.

There are various options to draw the dependent $\{\xi_t^*\}$; Shao (2010) proposes to draw these from a multivariate normal distributions, where the covariance between ξ_s^* and ξ_t^* is determined by a kernel function with as input the scaled distance $|s-t|/\ell$. The tuning parameter ℓ serves as a similar parameter as the block length in the MBB; the larger it is, the more serial dependence is captured. Smeekes and Urbain (2014b) and Friedrich et al. (2018) propose generating $\{\xi_t^*\}$ through an AR(1) process with normally distributed innovations and AR parameter γ, where γ is again a tuning parameter that determines how much serial dependence is captured. They label this approach the autoregressive wild bootstrap (AWB), and show that the AWB generally performs at least as well as Shao (2010) DWB in simulations.

Finally, one might consider the sieve wild bootstrap used in Cavaliere and Taylor (2009) and Smeekes and Taylor (2012), where the series $\{\widehat{\boldsymbol{u}}_t\}$ are first filtered through individual AR processes, and the wild bootstrap is applied afterwards to the residuals. However, as Smeekes and Urbain (2014c) show that this method cannot capture complex dynamic dependencies across series, it should not be used in this multivariate context. If common factors are believed to be the primary source of dependence across series, factor bootstrap methods such as those considered by Trapani (2013) or Gonçalves and Perron (2014) could be used as well.

5.4 High-Dimensional Cointegration

In this section, we discuss various recently proposed methods to model high-dimensional (co)integrated datasets. Similar to the high-dimensional modelling of stationary datasets, two main modelling approaches can be distinguished. One approach is to summarize the complete data into a much smaller and more manageable set through the extraction of common factors and their associated loadings, thereby casting the problem into the framework represented by (5.2.5). Another approach is to consider direct estimation of a system that is fully specified on the observable data as in (5.2.3), under the implicit assumption that the true DGP governing the long- and short-run dynamics is sparse, i.e. the number of non-zero coefficients in said relationships is small. These two approaches, however, rely on fundamentally different philosophies and estimation procedures, which constitute the topic of this section.[6]

[6]Some recent papers such as Onatski and Wang (2018) and Zhang et al. (2019b) have taken different, novel approaches to high-dimensional cointegration analysis. However, these methods do not directly lend themselves to forecasting and are therefore not discussed in this chapter.

5.4.1 Modelling Cointegration through Factor Structures

In this section, we discuss factor-based modelling of cointegrated datasets. Factor models are based on the intuitive notion that all variables in an economic system are driven by a small number of common shocks, which are often thought of as representing broad economic phenomena such as the unobserved business cycle. On (transformed) stationary macroeconomic datasets, the extracted factors have been successfully applied for the purpose of forecasting by incorporating them in dynamic factor models (Forni et al., 2005b), factor-augmented vector autoregressive (FAVAR) models (Bernanke et al., 2005a) or single-equation models (Stock and Watson, 2002a,b). Recent proposals are brought forward in the literature that allow for application of these techniques on non-stationary and possibly cointegrated datasets. We sequentially discuss the dynamic factor model proposed by Barigozzi et al. (2017, 2018) and the factor-augmented error correction model by Banerjee et al. (2014b, 2016). As both approaches require an a priori choice on the number of common factors, we follow the discussion with some remarks on the estimation of the factor dimension.

Dynamic factor models

A popular starting point for econometric modelling involving common shocks is the specification of a dynamic factor model. Recall our representation of an individual time series by

$$z_{i,t} = \mu_i + \tau_i t + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + u_{i,t}, \tag{5.4.1}$$

where $\boldsymbol{f}_t$ contains the $N - r$ common factors. Given a set of estimates for the unobserved factors, say $\hat{\boldsymbol{f}}_t$ for $t = 1, \dots, T$, one may directly obtain estimates for the remaining parameters in (5.4.1) by solving the least-squares regression problem[7]

$$\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\Lambda}}\right) = \underset{\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\Lambda}}{\arg\min} \sum_{t=1}^{T} \left(\boldsymbol{z}_t - \boldsymbol{\mu} - \boldsymbol{\tau} \boldsymbol{t} - \boldsymbol{\Lambda} \hat{\boldsymbol{f}}_t\right)^2. \tag{5.4.2}$$

The forecast for the realization of an observable time series at time period $T + h$ can then be constructed as

$$\hat{z}_{i,T+h|T} = \hat{\mu}_i + \hat{\tau}_i (T + h) + \hat{\boldsymbol{\lambda}}_i' \hat{\boldsymbol{f}}_{T+h|T}. \tag{5.4.3}$$

[7] Typically, the estimation procedure for $\hat{\boldsymbol{f}}_t$ provides the estimates $\hat{\boldsymbol{\Lambda}}$ as well, such that only the coefficients regulating the deterministic specification ought to be estimated.

This, however, requires the additional estimate $\hat{\boldsymbol{f}}_{T+h|T}$, which may be obtained through an explicit dynamic specification of the factors.

Barigozzi et al. (2018) assume that the differenced factors admit a reduced-rank vector autoregressive (VAR) representation, given by

$$\boldsymbol{S}(L)\Delta\boldsymbol{f}_t = \boldsymbol{C}(L)\boldsymbol{\nu}_t, \tag{5.4.4}$$

where $\boldsymbol{S}(L)$ is an invertible $((N-r)\times(N-r))$ matrix polynomial and $\boldsymbol{C}(L)$ is a finite degree $((N-r)\times q)$ matrix polynomial. Furthermore, $\boldsymbol{\nu}_t$ is a $(q\times 1)$ vector of white noise common shocks with $N-r>q$. Inverting the left-hand side matrix polynomial and summing both sides, gives rise to the specification

$$\begin{aligned}\boldsymbol{f}_t &= \boldsymbol{S}^{-1}(L)\boldsymbol{C}(L)\sum_{s=1}^{t}\boldsymbol{\nu}_s = \boldsymbol{U}(L)\sum_{s=1}^{t}\boldsymbol{\nu}_s \\ &= \boldsymbol{U}(1)\sum_{s=1}^{t}\boldsymbol{\nu}_s + \boldsymbol{U}^*(L)(\boldsymbol{u}_t-\boldsymbol{u}_0),\end{aligned} \tag{5.4.5}$$

where the last equation follows from application of the Beveridge-Nelson decomposition to $\boldsymbol{U}(L)=\boldsymbol{U}(1)+\boldsymbol{U}^*(L)(1-L)$. Thus, (5.4.5) reveals that the factors are driven by a set of common trends and stationary linear processes. Crucially, the assumption that the number of common shocks is strictly smaller than the number of integrated factors, i.e. $\boldsymbol{f}_t$ is a singular stochastic vector, implies that $\operatorname{rank}(\boldsymbol{U}(1)) = q-d$ for $0\leq d<q$. Consequently, there exists a full column rank matrix $\boldsymbol{B}_f$ of dimension $((N-r)\times(N-r-q+d))$ with the property that $\boldsymbol{B}_f'\boldsymbol{f}_t$ is stationary. Then, under the general assumption that the entries of $\boldsymbol{U}(L)$ are rational functions of L, Barigozzi et al. (2017) show that $\boldsymbol{f}_t$ admits a VECM representation of the form

$$\Delta\boldsymbol{f}_t = \boldsymbol{A}_f\boldsymbol{B}_f'\boldsymbol{f}_{t-1} + \sum_{j=1}^{p}\boldsymbol{G}_j\Delta\boldsymbol{f}_{t-j} + \boldsymbol{K}\boldsymbol{\nu}_t, \tag{5.4.6}$$

where $\boldsymbol{K}$ is a constant matrix of dimension $N-r\times q$.

Since the factors in (5.4.6) are unobserved, estimation of the system requires the use of a consistent estimate of the space spanned by $\boldsymbol{f}_t$. Allowing idiosyncratic components $\nu_{i,t}$ in (5.4.1) to be either $I(1)$ or $I(0)$, and allowing for the presence of a non-zero constant μ_i and linear trend τ_i, Barigozzi et al. (2018) propose an intuitive procedure that enables estimation of the factor space by the method of principal

components. First, the data is de-trended with the use of a regression estimate:

$$\tilde{z}_{i,t} = z_{i,t} - \hat{\tau}_i t,$$

where $\hat{\tau}_i$ is the OLS estimator of the trend in the regression of $z_{i,t}$ on an intercept and linear trend. Then, similar to the procedure originally proposed by Bai and Ng (2004), the factor loadings are estimated as $\hat{\boldsymbol{\Lambda}} = \sqrt{N}\hat{\boldsymbol{W}}$, where $\hat{\boldsymbol{W}}$ is the $(N \times (N-r))$ matrix with normalized right eigenvectors of $T^{-1}\sum_{t=1}^{T}\Delta\tilde{\boldsymbol{z}}_t\Delta\tilde{\boldsymbol{z}}_t'$ corresponding to the $N-r$ largest eigenvalues. The estimates for the factors are given by $\hat{\boldsymbol{f}}_t = \frac{1}{N}\hat{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{z}}_t$.

Plugging $\hat{\boldsymbol{f}}_t$ into (5.4.6) results in

$$\Delta\hat{\boldsymbol{f}}_t = \boldsymbol{A}_f\boldsymbol{B}_f'\hat{\boldsymbol{f}}_{t-1} + \sum_{j=1}^{p}\boldsymbol{G}_j\Delta\hat{\boldsymbol{f}}_{t-j} + \hat{\boldsymbol{\nu}}_t, \tag{5.4.7}$$

which can be estimated using standard approaches, such as the maximum likelihood procedure proposed by Johansen (1995a). Afterwards, the iterated one-step-ahead forecasts $\Delta\hat{\boldsymbol{f}}_{T+1|T}, \ldots, \Delta\hat{\boldsymbol{f}}_{T+h|T}$ are calculated from the estimated system, based on which the desired forecast $\hat{\boldsymbol{f}}_{T+h|T} = \hat{\boldsymbol{f}}_T + \sum_{k=1}^{h}\Delta\hat{\boldsymbol{f}}_{T+k|T}$ is obtained. The final forecast for $\hat{z}_{i,T+h|T}$ is then easily derived from (5.4.3).

Remark 5.3. Since the idiosyncratic components are allowed to be serially dependent or even $I(1)$, a possible extension is to explicitly model these dynamics. As a simple example, each $u_{i,t}$ could be modelled with a simple autoregressive model, from which the prediction $\hat{u}_{i,T+h|T}$ can be obtained following standard procedures (e.g. Hamilton, 1994, Ch. 4). This prediction is then added to (5.4.3), leading to the final forecast

$$\hat{z}_{i,T+h|T} = \hat{\mu}_i + \hat{\tau}_i(T+h) + \hat{\boldsymbol{\lambda}}_i\hat{\boldsymbol{f}}_{T+h|T} + \hat{u}_{i,T+h|T}.$$

This extension leads to substantial improvements in forecast performance in the macroeconomic forecast application presented in Section 5.5.

Factor-augmented error correction model

It frequently occurs that the variables of direct interest constitute only a small subset of the collection of observed variables. In this scenario, Banerjee, Marcellino, and Masten (2014b, 2016, 2017), henceforth referred to as BMM, propose to model only the series of interest in a VECM system, while including factors extracted from the full dataset to proxy for the missing information from the excluded observed time series.

The approach of BMM can be motivated starting from the common trend representation in (5.2.4). Partition the observed time series $\boldsymbol{z}_t = (\boldsymbol{z}'_{A,t}, \boldsymbol{z}'_{B,t})'$, where $\boldsymbol{z}_{A,t}$ is an $N_A \times 1$ vector containing the variables of interest. Then, we may rewrite (5.2.4) as

$$\begin{bmatrix} \boldsymbol{z}_{A,t} \\ \boldsymbol{z}_{B,t} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{bmatrix} + \begin{bmatrix} \boldsymbol{\tau}_A \\ \boldsymbol{\tau}_B \end{bmatrix} t + \begin{bmatrix} \boldsymbol{\Lambda}_A \\ \boldsymbol{\Lambda}_B \end{bmatrix} \boldsymbol{f}_t + \begin{bmatrix} \boldsymbol{u}_{A,t} \\ \boldsymbol{u}_{B,t} \end{bmatrix} \tag{5.4.8}$$

The idiosyncratic components in (5.4.8) are assumed to be $I(0)$.[8] Furthermore, both non-stationary $I(1)$ factors and stationary factors are admitted in the above representation. Contrary to Barigozzi et al. (2017), BMM do not require the factors in (5.4.8) to be singular.

To derive a dynamic representation better suited to forecasting the variables of interest, Banerjee et al. (2014b, 2017) use the fact that when the subset of variables is of a larger dimension than the factors, i.e. $N_A > N - r$, $\boldsymbol{z}_{A,t}$ and $\boldsymbol{f}_t$ cointegrate. As a result, the Granger Representation Theorem implies the existence of an error correction representation of the form

$$\begin{bmatrix} \Delta \boldsymbol{z}_{A,t} \\ \boldsymbol{f}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_f \end{bmatrix} + \begin{bmatrix} \boldsymbol{\tau}_A \\ \boldsymbol{\tau}_f \end{bmatrix} t + \begin{bmatrix} \boldsymbol{A}_A \\ \boldsymbol{A}_B \end{bmatrix} \boldsymbol{B}' \begin{bmatrix} \boldsymbol{z}_{A,t-1} \\ \boldsymbol{f}_{t-1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{e}_{A,t} \\ \boldsymbol{e}_{f,t} \end{bmatrix}. \tag{5.4.9}$$

To account for serial dependence in (5.4.9), Banerjee et al. (2014b) propose the approximating model

$$\begin{bmatrix} \Delta \boldsymbol{z}_{A,t} \\ \boldsymbol{f}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_f \end{bmatrix} + \begin{bmatrix} \boldsymbol{\tau}_A \\ \boldsymbol{\tau}_f \end{bmatrix} t + \begin{bmatrix} \boldsymbol{A}_A \\ \boldsymbol{A}_B \end{bmatrix} \boldsymbol{B}' \begin{bmatrix} \boldsymbol{z}_{A,t-1} \\ \boldsymbol{f}_{t-1} \end{bmatrix} + \sum_{j=1}^{p} \boldsymbol{\Phi}_j \begin{bmatrix} \Delta \boldsymbol{z}_{A,t-j} \\ \Delta \boldsymbol{f}_{t-j} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_{A,t} \\ \boldsymbol{\epsilon}_{f,t} \end{bmatrix}, \tag{5.4.10}$$

where the errors $\left(\boldsymbol{\epsilon}'_{A,t}, \boldsymbol{\epsilon}'_{f,t}\right)'$ are assumed i.i.d.

Similar to the case of the dynamic factor model in Section 5.4.1, the factors in the approximating model (5.4.10) are unobserved and need to be replaced with their corresponding estimates $\hat{\boldsymbol{f}}_t$. Under a set of mild assumptions, Bai (2004) shows that the space spanned by $\boldsymbol{f}_t$ can be consistently estimated using the method of principal components applied to the levels of the data. Assume that $\boldsymbol{f}_t = \left(\boldsymbol{f}'_{ns,t}, \boldsymbol{f}'_{s,t}\right)'$ where $\boldsymbol{f}_{ns,t}$ and $\boldsymbol{f}_{s,t}$ contain r_{ns} non-stationary and r_s stationary factors, respectively. Let $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T)$ be the $(N \times T)$ matrix of observed time series. Then,

[8] In principle, the proposed estimation procedure remains feasible in the presence of $I(1)$ idiosyncratic components. The theoretical motivation, however, relies on the concept of cointegration between the observable time series and a set of common factors. This only occurs when the idiosyncratic components are stationary.

Bai (2004) shows that $\boldsymbol{f}_{ns,t}$ is consistently estimated by $\hat{\boldsymbol{f}}_{ns,t}$, representing the eigenvectors corresponding to the r_{ns} largest eigenvalues of $\boldsymbol{Z}'\boldsymbol{Z}$, normalized such that $\frac{1}{T^2}\sum_{t=1}^{T}\hat{\boldsymbol{f}}_{ns,t}\hat{\boldsymbol{f}}'_{ns,t} = \boldsymbol{I}$. Similarly, $\boldsymbol{f}_{s,t}$ is consistently estimated by $\hat{\boldsymbol{f}}_{s,t}$, representing the eigenvectors corresponding to the next r_s largest eigenvalues of $\boldsymbol{Z}'\boldsymbol{Z}$, normalized such that $\frac{1}{T}\sum_{t=1}^{T}\hat{\boldsymbol{f}}_{s,t}\hat{\boldsymbol{f}}'_{s,t} = \boldsymbol{I}$.

The final step in the forecast exercise consists of plugging in $\hat{\boldsymbol{f}}_t = \left(\hat{\boldsymbol{f}}'_{ns,t}, \hat{\boldsymbol{f}}'_{s,t}\right)'$ into (5.4.10), leading to

$$\begin{bmatrix}\Delta \boldsymbol{z}_{A,t}\\ \hat{\boldsymbol{f}}_t\end{bmatrix} = \begin{bmatrix}\boldsymbol{\mu}_A\\ \boldsymbol{\mu}_f\end{bmatrix} + \begin{bmatrix}\boldsymbol{\tau}_A\\ \boldsymbol{\tau}_f\end{bmatrix} t + \begin{bmatrix}\boldsymbol{A}_A\\ \boldsymbol{A}_B\end{bmatrix} \boldsymbol{B}' \begin{bmatrix}\boldsymbol{z}_{A,t-1}\\ \hat{\boldsymbol{f}}_{t-1}\end{bmatrix} + \sum_{j=1}^{p} \boldsymbol{\Phi}_j \begin{bmatrix}\Delta \boldsymbol{z}_{A,t-j}\\ \Delta\hat{\boldsymbol{f}}_{t-j}\end{bmatrix} + \begin{bmatrix}\boldsymbol{\epsilon}_{A,t}\\ \boldsymbol{\epsilon}_{f,t}\end{bmatrix}. \quad (5.4.11)$$

Since in typical macroeconomic applications the number of factors is relatively small, feasible estimates for (5.4.11) can be obtained from the maximum likelihood procedure of Johansen (1995a). The iterated one-step-ahead forecasts $\Delta\hat{\boldsymbol{z}}_{A,T+1|T}, \ldots, \Delta\hat{\boldsymbol{z}}_{A,T+h|T}$ are calculated from the estimated system, which are then integrated to obtain the desired forecast $\hat{\boldsymbol{z}}_{A,T+h|T}$.

Estimating the number of factors

Implementation of the factor models discussed in this section requires an a priori choice regarding the number of factors. A wide variety of methods to estimate the dimension of the factors is available. The dynamic factor model of Barigozzi et al. (2017, 2018) adopts the estimation strategy proposed by Bai and Ng (2004), which relies on first-differencing the data. Since, under the assumed absence of $I(2)$ variables, all variables in this transformed dataset are stationary, the standard tools to determine the number of factors in the stationary setting are applicable. A non-exhaustive list is given by Bai and Ng (2002), Hallin and Liška (2007), Alessi et al. (2010), Onatski (2010) and Ahn and Horenstein (2013).

The factor-augmented error correction model of Banerjee et al. (2014b, 2016) adopts the estimation strategy proposed by Bai and Ng (2004), which extracts the factors from the data in levels. While the number of factors may still be determined based on the differenced dataset, Bai (2004) proposes a set of information criteria that allows for estimation of the number of non-stationary factors without differencing the data.

Conveniently, it is possible to combine factor selection procedures to separately determine the number of non-stationary and stationary factors. For example, the total number of factors, say $r_{ns} + r_s$, can be found based on the differenced dataset

and one of the information criteria in Bai and Ng (2002). Afterwards, the number of non-stationary factors, r_{ns}, is determined based on the data in levels using one of the the criteria from Bai (2004). The number of stationary factors follows from the difference between the two criteria. Recently, Barigozzi and Trapani (2018) propose a novel approach to discern the number of $I(0)$ factors, zero-mean $I(1)$ factors, and factors with a linear trend. Their method however requires that all idiosyncratic components are $I(0)$.

5.4.2 Sparse Models

Rather than extracting common factors, an alternative approach to forecasting with macroeconomic data is full-system estimation with the use of shrinkage estimators (e.g. De Mol et al., 2008; Stock and Watson, 2012; Callot and Kock, 2014). The general premise of shrinkage estimators is the so-called bias-variance trade-off, i.e. the idea that, by allowing a relatively small amount of bias in the estimation procedure, a larger reduction in variance may be attained. A number of shrinkage estimators, among which the lasso originally proposed by Tibshirani (1996), simultaneously perform variable selection and model estimation. Such methods are natural considerations when it is believed that the data generating process is sparse, i.e. only a small subset of variables among the candidate set is responsible for the variation in the variables of interest. Obviously, such a viewpoint is in sharp contrast with the philosophy underlying the common factor framework. However, in Chapter 2 it was demonstrated that even in cases where a sparse data generating process is deemed unrealistic, shrinkage estimators can remain attractive due to their aforementioned bias-variance trade-off.

For expositional convencience, we assume in this section that either $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ are zero or that $\boldsymbol{z}_t$ is de-meaned and de-trended. Defining $\boldsymbol{\Pi} = \boldsymbol{A}\boldsymbol{B}'$, model (5.2.3) is then given by

$$\Delta \boldsymbol{z}_t = \boldsymbol{\Pi}\boldsymbol{z}_{t-1} + \sum_{j=1}^{p} \boldsymbol{\Phi}_j \Delta \boldsymbol{z}_{t-j} + \boldsymbol{\epsilon}_t,$$

which in matrix notation reads as

$$\Delta \boldsymbol{Z} = \boldsymbol{\Pi}\boldsymbol{Z}_{-1} + \boldsymbol{\Phi}\Delta \boldsymbol{X} + \boldsymbol{E}, \tag{5.4.12}$$

where $\Delta \boldsymbol{Z} = (\Delta \boldsymbol{z}_1, \ldots, \Delta \boldsymbol{z}_T)$, $\boldsymbol{Z}_{-1} = (\boldsymbol{z}_0, \ldots, \boldsymbol{z}_{T-1})$, $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p)$ and $\Delta \boldsymbol{X} = (\Delta \boldsymbol{x}_0, \ldots, \Delta \boldsymbol{x}_{T-1})$, with $\boldsymbol{x}_t = \left(\boldsymbol{z}_t', \ldots, \boldsymbol{z}_{t-p+1}'\right)'$.

Full-system estimation

Several proposals to estimate (5.4.12) with the use of shrinkage estimators are brought forward in recent literature. Liao and Phillips (2015) proposes an automated approach that simultaneously enables sparse estimation of the coefficient matrices $(\boldsymbol{\Pi}, \boldsymbol{\Phi})$, including the cointegrating rank of $\boldsymbol{\Pi}$ and the short-run dynamic lag order in $\boldsymbol{\Phi}$. However, while the method has attractive (fixed-dimensional) theoretical properties, the estimation procedure involves non-standard optimization over the complex plane and is difficult to implement even in low dimensions, as also noted by Liang and Schienle (2019). Accordingly, we do not further elaborate on their proposed method, but refer the interested reader to the original paper.

Liang and Schienle (2019) develop an automated estimation procedure that makes use of a QR-decomposition of the long-run coefficient matrix. They propose to first regress out the short-run dynamics, by post-multiplying (5.4.12) with $\boldsymbol{M} = \boldsymbol{I}_T - \Delta\boldsymbol{X}' \left(\Delta\boldsymbol{X}'\Delta\boldsymbol{X}\right)^{-1} \Delta\boldsymbol{X}$, resulting in

$$\Delta\tilde{\boldsymbol{Z}} = \boldsymbol{\Pi}\tilde{\boldsymbol{Z}}_{-1} + \tilde{\boldsymbol{E}}, \tag{5.4.13}$$

with $\Delta\tilde{\boldsymbol{Z}} = \Delta\boldsymbol{Z}M$, $\tilde{\boldsymbol{Z}}_{-1} = \boldsymbol{Z}_{-1}M$ and $\tilde{\boldsymbol{E}} = \boldsymbol{E}M$. The key idea behind the method proposed by Liang and Schienle is to decompose the long-run coefficient matrix into

$$\boldsymbol{\Pi}' = \boldsymbol{Q}\boldsymbol{R},$$

where $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_N$ and $\boldsymbol{R}$ is an upper-triangular matrix. Such a representation can be be calculated from the QR-decomposition of $\boldsymbol{\Pi}$ with column pivoting.

The column pivoting orders the columns in $\boldsymbol{R}$ according to size, such that zero elements occur at the ends of the rows. As a result, the rank of $\boldsymbol{\Pi}$ corresponds to the number of non-zero columns in $\boldsymbol{R}$. Exploiting this rank property requires an initial estimator for the long-run coefficient matrix, such as the OLS estimator

$$\hat{\boldsymbol{\Pi}}_{OLS} = \left(\Delta\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}'_{-1}\right)\left(\tilde{\boldsymbol{Z}}_{-1}\tilde{\boldsymbol{Z}}'_{-1}\right)^{-1},$$

proposed by Liang and Schienle (2019). The QR-decomposition with column-pivoting is then calculated from $\hat{\boldsymbol{\Pi}}'_{OLS}$, resulting in the representation $\hat{\boldsymbol{\Pi}}_{OLS} = \hat{\boldsymbol{R}}'_{OLS}\hat{\boldsymbol{Q}}'_{OLS}$.[9] Since the unrestricted estimator $\hat{\boldsymbol{\Pi}}_{OLS}$ will be full-rank, $\hat{\boldsymbol{R}}_{OLS}$ is a full-rank matrix

[9] As part of their theoretical contributions, Liang and Schienle (2019) show that the first r columns of $\hat{\boldsymbol{Q}}$ consistently estimate the space spanned by the cointegrating vectors $\boldsymbol{B}$ in (5.2.3), in an asymptotic framework where the dimension N is allowed to grow at rate $T^{1/4-\nu}$ for $\nu > 0$.

as well. However, by the consistency of $\hat{\boldsymbol{\Pi}}_{OLS}$ and the ordering induced by the column-pivoting step, the last $N - r$ columns are expected to contain elements that are small in magnitude. Accordingly, a well-chosen shrinkage estimator that penalizes the columns of $\boldsymbol{R}$ may be able to separate the relevant from the irrelevant columns.

Let $\hat{\boldsymbol{R}} = (\hat{\boldsymbol{r}}_1, \dots, \hat{\boldsymbol{r}}_N)$, $\hat{\boldsymbol{r}}_j = (\hat{r}_{1,j}, \dots, \hat{r}_{N,j})'$, $\|\hat{\boldsymbol{r}}_j\|_2 = \sqrt{\sum_{i=1}^{N} \hat{r}_{i,j}^2}$ and $\hat{\mu}_k = \sqrt{\sum_{i=k}^{N} \hat{r}_{k,i}^2}$. Then, the estimator for $\boldsymbol{R}$ is defined as

$$\hat{\boldsymbol{R}} = \underset{\boldsymbol{R}}{\arg\min} \left\| \Delta \boldsymbol{Z} - \boldsymbol{R}' \hat{\boldsymbol{Q}}' \boldsymbol{Z}_{-1} \right\|_2^2 + \lambda \sum_{j=1}^{N} \frac{\|\hat{\boldsymbol{r}}_j\|_2}{\hat{\mu}_j}, \tag{5.4.14}$$

where λ is a tuning parameter that controls the degree of regularization, with larger values resulting in more shrinkage. Weighting the penalty for each group by $\hat{\mu}_j$ puts a relatively higher penalty on groups for which the initial OLS estimates are small. The estimator clearly penalizes a set of pre-defined groups of coefficients, i.e. the columns of $\boldsymbol{R}$, and, therefore, is a variant of the group lasso for which numerous algorithms are available (e.g. Meier et al., 2008; Friedman et al., 2010; Simon et al., 2013). The final estimate for the long-run coefficient matrix is obtained as $\hat{\boldsymbol{\Pi}} = \hat{\boldsymbol{R}}' \hat{\boldsymbol{Q}}'_{OLS}$.

The procedure detailed thus far focuses solely on estimation of the long-run relationships and requires an a priori choice of the lag order p. Furthermore, a necessary assumption is that initial OLS estimates are available, thereby restricting the admissible dimension of the system to $N(p+1) < T$. Within this restricted dimension, the short-run coefficient matrix $\boldsymbol{\Phi}$ can be consistently estimated by OLS and the corresponding lag order may be determined by standard information criteria such as the BIC. Alternatively, a second adaptive group lasso can be employed to obtain the regularized estimates $\hat{\boldsymbol{\Phi}} = \left(\hat{\boldsymbol{\Phi}}_1, \dots, \hat{\boldsymbol{\Phi}}_p\right)$, see Liang and Schienle (2019, p. 425) for details. The lag order is then determined by the number of non-zero matrices $\hat{\boldsymbol{\Phi}}_i$ for $i \in \{1, \dots, p\}$.

Wilms and Croux (2016) propose a penalized maximum likelihood estimator to estimate sparse VECMs. Instead of estimating the cointegrating rank and coefficient matrices for a fixed lag order, the method of Wilms and Croux enables joint estimation of the lag order and coefficient matrices for a given cointegrating rank. Additionally, the penalized maximum likelihood procedure does not require the availability of initial OLS estimates and, therefore, notwithstanding computational constraints, can be applied to datasets of arbitrary dimension. Under the assumption of multivariate normality of the errors, i.e. $\boldsymbol{\epsilon}_t \sim \mathbb{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, the penalized negative log-likelihood is

given by

$$\mathscr{L}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\Phi},\boldsymbol{\Omega}) = \frac{1}{T}\mathrm{tr}\left((\Delta\boldsymbol{Z} - \boldsymbol{A}\boldsymbol{B}'\boldsymbol{Z}_{-1} - \boldsymbol{\Phi}\Delta\boldsymbol{X})'\boldsymbol{\Omega}(\Delta\boldsymbol{Z} - \boldsymbol{A}\boldsymbol{B}'\boldsymbol{Z}_{-1} - \boldsymbol{\Phi}\Delta\boldsymbol{X})\right) - \log|\boldsymbol{\Omega}| + \lambda_1 P_1(\boldsymbol{B}) + \lambda_2 P_2(\boldsymbol{\Phi}) + \lambda_3 P_3(\boldsymbol{\Omega}), \tag{5.4.15}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, and P_1, P_2 and P_3 being three penalty functions. The cointegrating vectors, short-run dynamics, and covariance matrix are penalized as

$$P_1(\boldsymbol{B}) = \sum_{i=1}^{N}\sum_{j=1}^{r}|\beta|_{i,j}\,, \quad P_2(\boldsymbol{\Phi}) = \sum_{i=1}^{N}\sum_{j=1}^{Np}|\phi_{i,j}|\,, \quad P_3(\boldsymbol{\Omega}) = \sum_{i,j=1,i\neq j}^{N}|\omega_{i,j}|\,,$$

respectively. The use of L_1-penalization enables some elements to be estimated as exactly zero. The solution that minimizes (5.4.15) is obtained through an iterative updating scheme, where the solution for a coefficient matrix is obtained by minimizing the objective function conditional on the remaining coefficient matrices. The full algorithm is described in detail in Wilms and Croux (2016, p. 1527-1528) and R code is provided by the authors online.[10]

Single-equation estimation

Frequently, the forecast exercise is aimed at forecasting a small number of time series based on a large number of potentially relevant variables. The means of data reduction thus far considered utilize either data aggregation or subset selection. However, in cases where the set of target variables is small, a substantial reduction in dimension can be obtained through the choice of appropriate single-equation representations for each variable separately.

In Chapter 3, we first propose the Penalized Error Correction Selector (SPECS) as an automated single-equation modelling procedure on high-dimensional (co)integrated datasets. For the sake of completion, we briefly recollect its main features here. Assume that the N-dimensional observed time series admits the decomposition $\boldsymbol{z}_t = (y_t, \boldsymbol{x}_t')'$, where y_t is the variable of interest and $\boldsymbol{x}_t$ are variables that are considered as potentially relevant in explaining the variation in y_t. Starting from the VECM system (5.4.12), a single-equation representation for Δy_t can be obtained by conditioning on the contemporaneous differences $\Delta\boldsymbol{x}_t$. This results in

$$\Delta y_t = \boldsymbol{\delta}'\boldsymbol{z}_{t-1} + \boldsymbol{\pi}'\boldsymbol{w}_t + \epsilon_{y,t}, \tag{5.4.16}$$

[10] https://feb.kuleuven.be/public/u0070413/SparseCointegration/

where $\boldsymbol{w}_t = (\Delta \boldsymbol{x}_t', \Delta \boldsymbol{z}_{t-1}', \ldots, \Delta \boldsymbol{z}_{t-p}')'$[11]. The number of parameters to be estimated in the single-equation model (5.4.16) is $2N(p+2)-1$ as opposed to the original $2Nr + N^2p$ parameters in (5.4.12). Nonetheless, for large N the total number of parameters may still be too large to estimate precisely by ordinary least squares, if possible at all. Therefore, we propose a shrinkage procedure defined as

$$\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\pi}} = \underset{\boldsymbol{\delta}, \boldsymbol{\pi}}{\arg\min} \sum_{t=1}^{T} \left(\Delta y_t - \boldsymbol{\delta}' \boldsymbol{z}_{t-1} + \boldsymbol{\pi}' \boldsymbol{w}_t\right)^2 + P_\lambda(\boldsymbol{\delta}, \boldsymbol{\pi}). \tag{5.4.17}$$

The penalty function takes on the form

$$P_\lambda(\boldsymbol{\delta}, \boldsymbol{\pi}) = \lambda_G \|\boldsymbol{\delta}\| + \lambda_\delta \sum_{i=1}^{N} \omega_{\delta,i}^{k_\delta} |\delta_i| + \lambda_\pi \sum_{j=1}^{N(p+1)-1} \omega_{\pi,j}^{k_\pi} |\pi_j|, \tag{5.4.18}$$

where $\omega_{\delta,i}^{k_\delta} = 1/\left|\hat{\delta}_{Init,i}\right|^{k_\delta}$ and $\omega_{\pi,j}^{k_\pi} = 1/|\hat{\pi}_{Init,j}|^{k_\pi}$, with $\hat{\boldsymbol{\delta}}_{Init}$ and $\hat{\boldsymbol{\pi}}_{Init}$ being some consistent initial estimates, such as OLS or ridge estimates. The tuning parameters k_δ and k_π regulate the degree to which the initial estimates affect the penalty weights.

SPECS simultaneously employs individual penalties on all coefficients and a group penalty on $\boldsymbol{\delta}$, the implied cointegrating vector. Absent of cointegration, this cointegrating vector is equal to zero, in which case the group penalty promotes the removal of the lagged levels as a group.[12] In the presence of cointegration, however, the implied cointegrating vector may still contain many zero elements. The addition of the individual penalties allow for correct recovery of this sparsity pattern. This combination of penalties is commonly referred to as the sparse group lasso and R code is provided by the author of this thesis.[13]

In the single-equation model, the variation in y_t is explained by contemporaneous realizations of the conditioning variables $\boldsymbol{x}_t$. Therefore, forecasting the variable of interest requires forecasts for the latter as well, unless their realizations become available to the researcher prior to the realizations of y_t. SPECS is therefore highly suited to nowcasting applications. While not originally developed for the purpose of forecasting, direct forecasts with SPECS can be obtained by modifying the objective

[11] Details regarding the relationship between the components of the single-equation model (5.4.16) and the full system (5.2.3) are provided in Chapter 3.

[12] As argued in Chapter 3, the group penalty is not formally required for consistent selection and estimation of the non-zero coefficients.

[13] `https://sites.google.com/view/etiennewijler/code?authuser=0`

function to

$$\sum_{t=1}^{T} \left(\Delta_h y_t - \boldsymbol{\delta}' \boldsymbol{z}_{t-1} + \boldsymbol{\pi}' \boldsymbol{w}_t\right)^2 + P_\lambda(\boldsymbol{\delta}, \boldsymbol{\pi}),$$

where $\Delta_h y_t = y_{t+h} - y_t$. The direct h-step ahead forecast is then simply obtained as $\hat{y}_{T+h|T} = y_T + \hat{\boldsymbol{\delta}}' \boldsymbol{z}_{T-1} + \hat{\boldsymbol{\pi}}' \boldsymbol{w}_T$.

5.5 Empirical Applications

In this section we evaluate the methods discussed in Sections 5.3 and 5.4 in two empirical applications. First we forecast several US macroeconomic variables using the FRED-MD dataset of McCracken and Ng (2016). The FRED-MD dataset is a well-established and popular source for macroeconomic forecasting, and allows us to evaluate the methods in an almost controlled environment. Second we consider nowcasting Dutch unemployment using Google Trends data on frequencies of unemployment-related queries. This application not only highlights the potential of novel high-dimensional datasets for macroeconomic purposes, but also puts the methods to the test in a more difficult environment where less theoretical guidance is available on the properties of the data.

5.5.1 Macroeconomic Forecasting Using the FRED-MD Dataset

We consider forecasting eight US macroeconomic variables from the FRED-MD dataset at 1, 6 and 12 months forecast horizons, corresponding to the same variables considered in the empirical application of Chapter 2. We first focus on the strategy discussed in Section 5.3 where we first transform all series to I(0) before estimating the forecasting models. We illustrate the unit root testing methods, and show the empirical consequences of specification changes in the orders of integration. Next, we analyze the methods discussed in Section 5.4, and compare their forecast accuracy.

Transformations to stationarity

As the FRED-MD series have already been classified by McCracken and Ng (2016), we have a benchmark for our own classification using the unit root testing methodology discussed in Section 5.3. We consider the autoregressive wild bootstrap as described in Section 5.3.2 in combination with the union test in (5.3.1). We set the AWB parameter γ equal to 0.85, which implies that over a year of serial dependence is captured by the bootstrap. Lag lengths in the ADF regressions are selected by the rescaled MAIC

criterion of Cavaliere et al. (2015), which is robust to heteroskedasticity. To account for multiple testing, we control the false discovery rate at 5% using the bootstrap method of Romano et al. (2008a) (labelled as 'BFDR') and apply the sequential test procedure of Smeekes (2015) (labelled as 'BSQT') with a significance level of 5% and evenly spaced 0.05 quantiles such that $p_k = [0.05(k-1)]$ for $k = 1, \dots, 20$. We also perform the unit root tests on each series individually (labelled as 'iADF') with a significance level of 5%.

As some series in the FRED-MD are likely $I(2)$, we need to extend the methodology to detect these as well. We consider two ways to do so. First, we borrow information about the $I(2)$ series from the official FRED-MD classification, and take first differences of the series deemed to be $I(2)$. We then put these first differences together with the other series in levels and test for unit roots. This strategy ensures that the $I(2)$ series are classified at least as $I(1)$, and we only need to perform a single round of unit root testing. Our second approach is fully data-driven and follows a multivariate extension of the 'Pantula principle' (Pantula, 1989), where we first test for a unit root in the first difference of all series. The series for which the null cannot be rejected are classified as $I(2)$ and removed from the sample. The remaining series are then tested in levels and consequently classified as $I(1)$ or $I(0)$. In the results we append an acronym with a 1 if the first strategy is followed, and with a 2 if the second strategy is followed.[14]

As a final method, we include a 'naive' unit root testing approach that we believe is representative of casual unit root testing applied by many practitioners who, understandably, may not pay too much detailed attention to the unit root testing. In particular, we use the `adf.test` function from the popular R package 'tseries' (Trapletti and Hornik, 2018), and apply it with its default options, which implies performing individual ADF tests with a trend and setting a fixed lag length as a function of the sample size.[15] Our goal is not to discuss the merits of this particular unit root test procedure, but instead to highlight the consequences of casually using a 'standard' unit root test procedure that does not address the issues described in Section 5.3. Figure 5.1 presents the found orders of integration. Globally the classifi-

[14] We take logarithmic transformations of the series before differencing when indicated by the official FRED-MD classification. Determining when a logarithmic transformation is appropriate is a daunting task for such a high-dimensional system as it seems difficult to automatize, especially as it cannot be seen separately from the determination of the order of integration (Franses and McAleer, 1998; Kramer and Davies, 2002). Klaassen et al. (2017) propose a high-dimensional method to determine an appropriate transformation model, but it is not trivial how to combine their method with unit root testing. Therefore we apply the 'true' transformations such that we can abstract from this issue.

[15] The lag length is set equal to $\lfloor (T-1)^{1/3} \rfloor$.

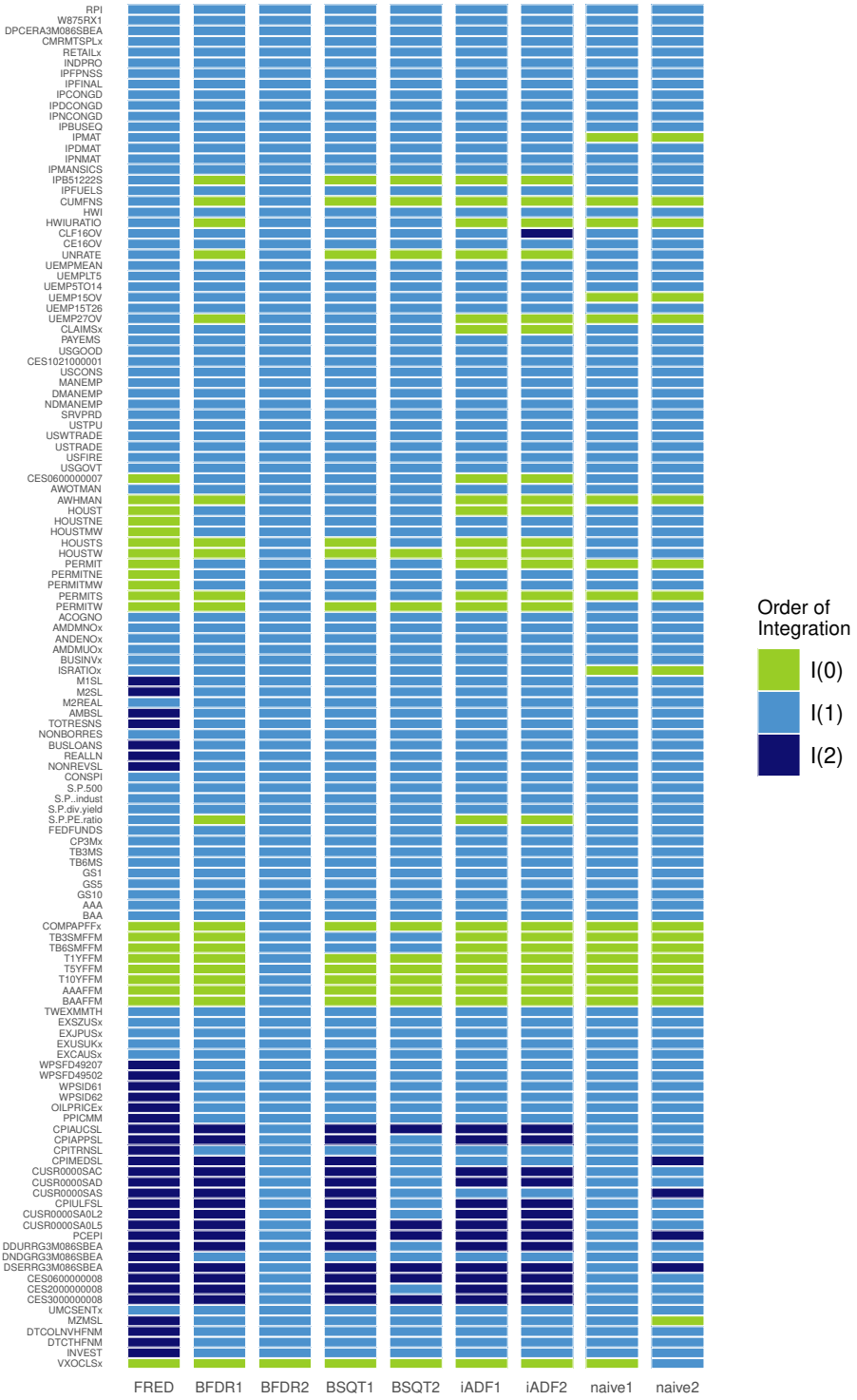

Figure 5.1: Classification of integration order of the FRED-MD dataset.

cation appears to agree among the different methods, which is comforting, although some important differences can be noted. First, none of the data-driven methods finds as many $I(2)$ series as the FRED classification does. Indeed, this may not be such a surprising result, as it remains a debated issue among practitioners whether these series should be modelled as $I(1)$ or as $I(2)$, see for example the discussion in Marcellino et al. (2006b).

Second, although most methods yield fairly similar classifications, the clear outlier is BFDR2, which finds all series but one to be $I(1)$. The FDR controlling algorithm may, by construction, be too conservative in the early stages of the algorithm when few rejections R have been recorded, yet too liberal in the final stages upon finding many rejections. Indeed, when testing the first differences of all series for a unit root, the FRED classification tells us that for most of the series the null can be rejected. When the algorithm arrives at the $I(2)$ series, the unit root hypothesis will already have been rejected for many series. With R being that large, the number of false rejections F can be relatively large too without increasing the FDR too much. Hence, incorrectly rejecting the null for the $I(2)$ series will fall within the 'margins of error' and thus lead to a complete rejection of all null hypotheses. In the second step the FDR algorithm then appears to get 'stuck' in the early stages, resulting in only a single rejection. This risk of the method getting stuck early on was also observed by Smeekes (2015) and can be explained by the fact that early on in the step-down procedure, when R is small, FDR is about as strict as FWE. It appears that in this case the inclusion of the $I(2)$ series in levels rather than differences is just enough to make the algorithm get stuck.

Third, even though iADF does not control for multiple testing, its results are fairly similar to BSQT and FDR1. It therefore appears explicitly controlling for multiple testing is not the most important in this application, and sensible unit root tests, even when applied individually, will give reasonable answers. On first glance even using the 'naive' strategy appears not be very harmful. However, upon more careful inspection of the results, we can see that it does differ from the other methods. In particular, almost no $I(2)$ series are detected by this strategy, and given that there is no reason to prefer it over the other methods, we recommend against its use.

Forecast comparison after transformations

While determining an appropriate order of integration may be of interest in itself, our goal here is to evaluate its impact on forecast accuracy. As such, we next evaluate if, and how, the chosen transformation impacts the actual forecast performance of the

BFDR, BSQT and iADF methods, all in both strategies considered, in comparison with the official FRED classification.

We forecast eight macroeconomic series in the FRED-MD dataset using data from July 1972 to October 2018. The series of interest consist of four real series, namely real production income (RPI), total industrial production (INDPRO), real manufacturing and trade industries sales (CMRMTSPLx) and non-agricultural employees (PAYEMS), and four nominal series, being the producer index for finished good (WPSFD49207), consumer price index - total (CPIAUCSL), consumer price index - less food (CPIULFSL) and the PCE price deflator (PCEPI). Each series is forecast h months ahead, where we consider the forecast horizons $h = 1, 6, 12$. All models are estimated on a rolling window spanning ten years, i.e. containing 120 observations. Within each window, we regress every time series on a constant and linear trend and obtain the corresponding residuals. For the stationary methods, these residuals are transformed to stationarity according to the results of the unit root testing procedure. Each model is fitted to these transformed residuals, after which the h-step ahead forecast is constructed as an iterated one-step-ahead forecast, when possible, and transformed to levels, if needed. The final forecast is obtained by adding the level forecast of the transformed residuals to the forecast of the deterministic components. We briefly describe the implementation of each method below.

We consider four methods here. The first method is a standard vector autoregressive (VAR) model, fit on the eight variables of interest. Considering only the eight series of interest, however, may result in a substantial loss of relevant information contained in the remaining variables in the complete dataset. Therefore, we also consider a factor-augmented vector autoregressive model (FAVAR) in the spirit of Bernanke et al. (2005a), which includes factors as proxies for this missing information. We extract four factors from the complete and transformed dataset and fit two separate FAVAR models containing these four factors, in addition to either the four real or the four nominal series. Rather than focusing on the estimation of heavily parameterized full systems, one may attempt to reduce the dimensionality by considering single-equation models, as discussed in Section 5.4.2. Conditioning the variable of interest on the remaining variables in the dataset, results in an autoregressive distributed lag model with $M = N(p+1) - 1$ parameters. For large N, shrinkage may still be desirable. Therefore, we include a penalized autoregressive distributed lag model (PADL) in the comparison, which is based on the minimization of

$$\sum_{t=1}^{T} \left(y_t^h - \boldsymbol{\pi}' \boldsymbol{w}_t\right)^2 + \lambda \sum_{j=1}^{M} \omega_{\pi,j}^{k_\pi} \left|\pi\right|_j , \tag{5.5.1}$$

where

$$y_t^h = \begin{cases} y_{t+h} - y_t & \text{if } y_t \sim I(1), \\ y_{t+h} - y_t - \Delta y_t & \text{if } y_t \sim I(2). \end{cases} \tag{5.5.2}$$

Furthermore, $\boldsymbol{w}_t$ contains contemporaneous values of all transformed time series except y_t, and three lags of all transformed time series. The weights $\omega_{\pi,j}^{k_\pi}$ are as defined in Section 5.4.2. In essence, this can be seen as an implementation of SPECS with the build-in restriction that $\boldsymbol{\delta} = \mathbf{0}$, thereby ignoring cointegration. Finally, the concept of using factors as proxies for missing information remains equally useful for single-equation models. Accordingly, we include a factor-augmented penalized autoregressive distributed model (FAPADL) which is a single-equation model derived from a FAVAR. We estimate eight factors on the complete dataset, which are added to the eight variables of interest in the single-equation model. This is then estimated in accordance to (5.5.1), with $\boldsymbol{w}_t$ now containing contemporaneous values and three lags of the eight time series of interest and the eight factors. The PADL and FAPADL are variants of the adaptive lasso and we implement these in R based on the popular 'glmnet' package (Friedman et al., 2010). The lag order for the VAR and FAVAR are chosen by the BIC criterion, with a maximum lag order of three.

Our goal is not to be exhaustive, but we believe these four methods cover a wide enough range of available high-dimensional forecast methods such that our results cannot be attributed to the choice of a particular forecasting method and instead genuinely reflect the effect of different transformations to stationarity. For the sake of space, we only report the results based on the FAVAR here for 1 month and 12 months ahead forecasts, as these are representative for the full set of results (which are available upon request). Generally, we find the same patterns within each method as we observe for the FAVAR, though they may be more or less pronounced. Overall the FAVAR is the most accurate of the four methods considered, which is why we choose to focus on it.

We compare the methods through their relative Mean Squared Forecast Errors (MSFEs), where the AR model is taken as benchmark. To attach a measure of statistical significance to these MSFEs, we obtain 90% Model Confidence Sets (MCS) of the best performing model. We obtain the MCS using the autoregressive wild bootstrap as in Chapter 2.

The results are given in Figures 5.2 and 5.3. For the one-month-ahead forecast the results are close for the different transformation methods, but for the twelve-months-ahead forecasts, we clearly see big differences for the nominal series. Inspection of

the classifications in Figure 5.1 shows that the decisive factor is the classification of the dependent variable. For the three price series, the methods that classify these as $I(1)$ rather than $I(2)$ obtain substantial gains in forecast accuracy. Interestingly, the FRED classification finds these series to be $I(2)$, and thus deviating from the official classification can lead to substantial gains. These results are in line with the results of Marcellino et al. (2006b), who also find that modelling price series as $I(1)$ rather than $I(2)$ results in better forecast accuracy.

As the outlying BFDR2 classification also classifies these series as I(1), this 'lucky shot' eclipses any losses from the missclassification of the other series. However, for the real series it can be observed that BFDR2 does indeed always perform somewhat worse than the other methods, although the MCS does not find it to be significant everywhere.

Concluding, missclassification of the order of integration can have an effect on the performance of high-dimensional forecasting methods. However, unless the dependent variable is miss-classified, the high-dimensional nature of the data also ensures that this effect is smoothed out. On the other hand, correct classification of the dependent variable appears to be crucial, in particular regarding the classification as $I(1)$ versus $I(2)$.

Forecast comparisons for cointegration methods

The forecast exercise for the methods that are able to take into account the cointegrating properties of the data proceeds along the same lines as in Section 5.5.1. A noteworthy exception is that the time series that are considered I(1) in the FRED-MD classification are now kept in levels, whereas those that are considered as I(2) are differenced once. The methods included in the comparison are: (i) the factor error correction model (FECM) by Banerjee et al. (2014b, 2016, 2017), (ii) the non-stationary dynamic factor model (N-DFM) by Barigozzi et al. (2017, 2018), (iii) the maximum-likelihood procedure (ML) by Johansen (1995a), (iv) the QR-decomposed VECM (QR-VECM) by Liang and Schienle (2019), (v) the penalized maximum-likelihood (PML) by Wilms and Croux (2016), (vi) the single-equation penalized error correction selector (SPECS) and (vii) a factor-augmented SPECS (FASPECS). The latter method is simply the single-equation model derived from the FECM, based on the same principles as the FAPADL from the previous section. It is worth noting that the majority of these non-stationary methods have natural counterparts in the stationary world; the ML procedure compares directly to the VAR model, FECM compares to FAVAR, and SPECS and FA-SPECS to PADL and FAPADL, respectively. Finally,

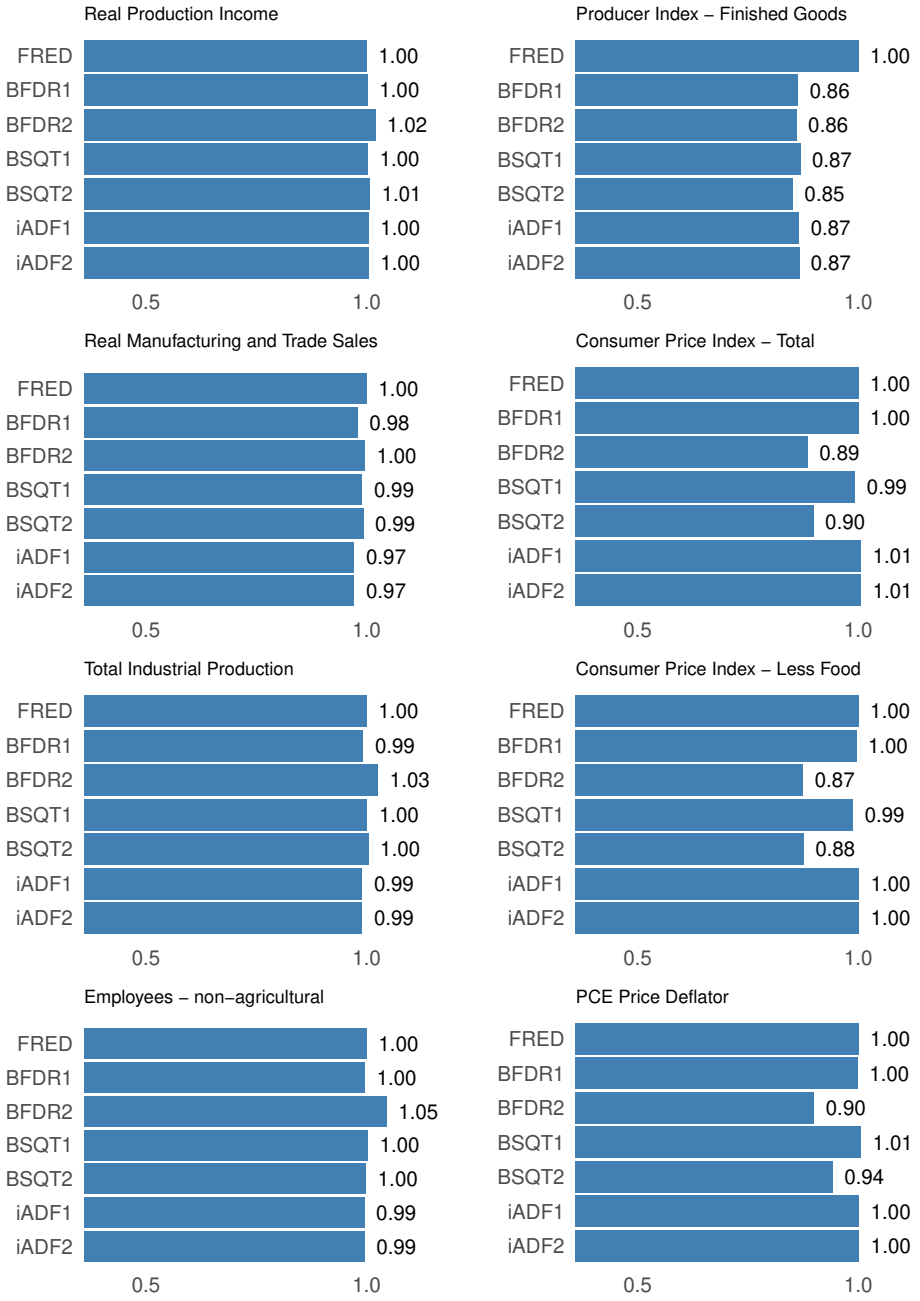

Figure 5.2: MCS and relative MSFEs for 1-month horizon. Methods that are included in the MCS are depicted as blue and methods that are excluded from the MCS are depicted in red.

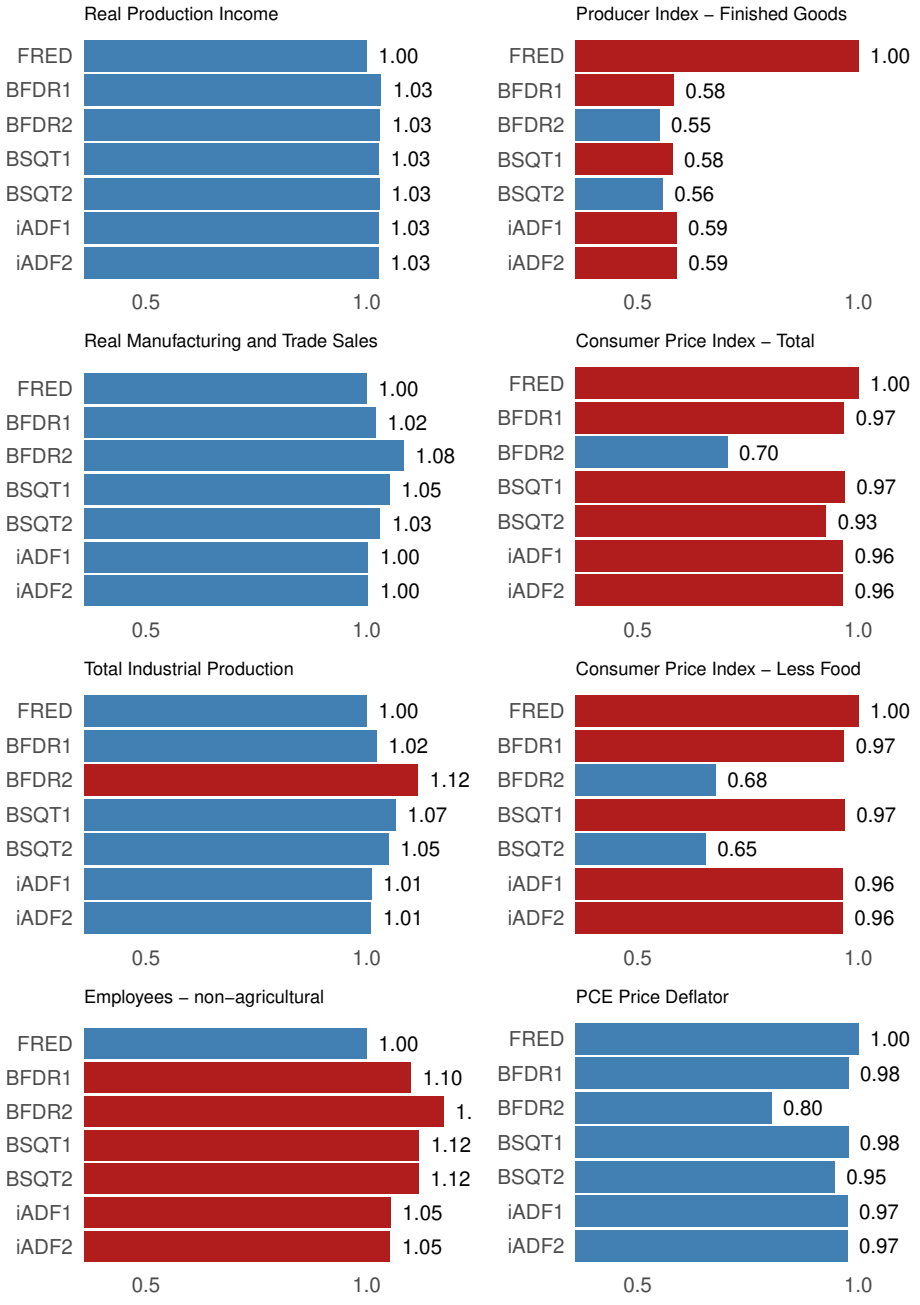

Figure 5.3: MCS and relative MSFEs for 12-month horizon. Methods that are included in the MCS are depicted as blue and methods that are excluded from the MCS are depicted in red.

all methods are compared against an AR model fit on the dependent variable, the latter being transformed according to the original FRED codes.

We briefly discuss some additional implementation choices for the non-stationary methods. For all procedures that require an estimate of the cointegrating rank, we use the information criteria proposed by Cheng and Phillips (2009). The only exception is the PML method, for which the cointegrating rank is determined by the procedure advocated in Wilms and Croux (2016). Similar to Banerjee et al. (2014b), we do not rely on information criteria to select the number of factors, but rather fix the number of factors in the implementation of the FECM and N-DFM methods to four.[16] In the N-DFM approach, we model the idiosyncratic components of the target variables as simple AR models. The ML procedure estimates a VECM system on the eight variables of interest. In congruence with the implementation of the stationary methods, the lag order for FECM, N-DFM and ML is chosen by the BIC criterion, with a maximum lag order of three. The QR-VECM and PML methods are estimated on a dataset containing the eight series of interest and an additional 17 variables, informally selected based on their unique information within each economic category. Details are provided in Table 5.1. We incorporate only a single lag in the QR-VECM implementation, necessitated by the requirement of initial OLS estimates. SPECS estimates the model

$$y_t^h = \boldsymbol{\delta}' \boldsymbol{z}_{t-1} + \boldsymbol{\pi}' \boldsymbol{w}_t + \epsilon_{y,t},$$

where y_t^h is defined in (5.5.2), with the order of integration based on the original FRED codes. Note that the variables included in $\boldsymbol{z}_t$ are either the complete set of 124 time series or the eight time series of interest plus an additional eight estimated factors, depending on whether the implementation concerns SPECS or FA-SPECS, respectively. Finally, all parameters that regulate the degree of shrinkage are chosen by time series cross-validation, proposed by Hyndman and Athanasopoulos (2018) and discussed in a context similar to the current analysis in Chapter 2.

Results are given in Figure 5.4-5.6. Considering first the 1-month ahead predictions, we observe similar forecasting performance on the first three real series (RPI,CMRMTSPLx, INDPRO) with almost none of the methods being excluded from the 90% model confidence set. The employment forecasts of the AR benchmark and the FAVAR approach are considered superior to those of the other methods. On

[16] In untabulated results, we find that the forecast performance does not improve when the number of factors is selected by the information criteria by Bai (2004). Neither does the addition of a stationary factor computed from the estimated idiosyncratic component, in the spirit of Banerjee et al. (2014b). Both strategies are therefore omitted from the analysis.

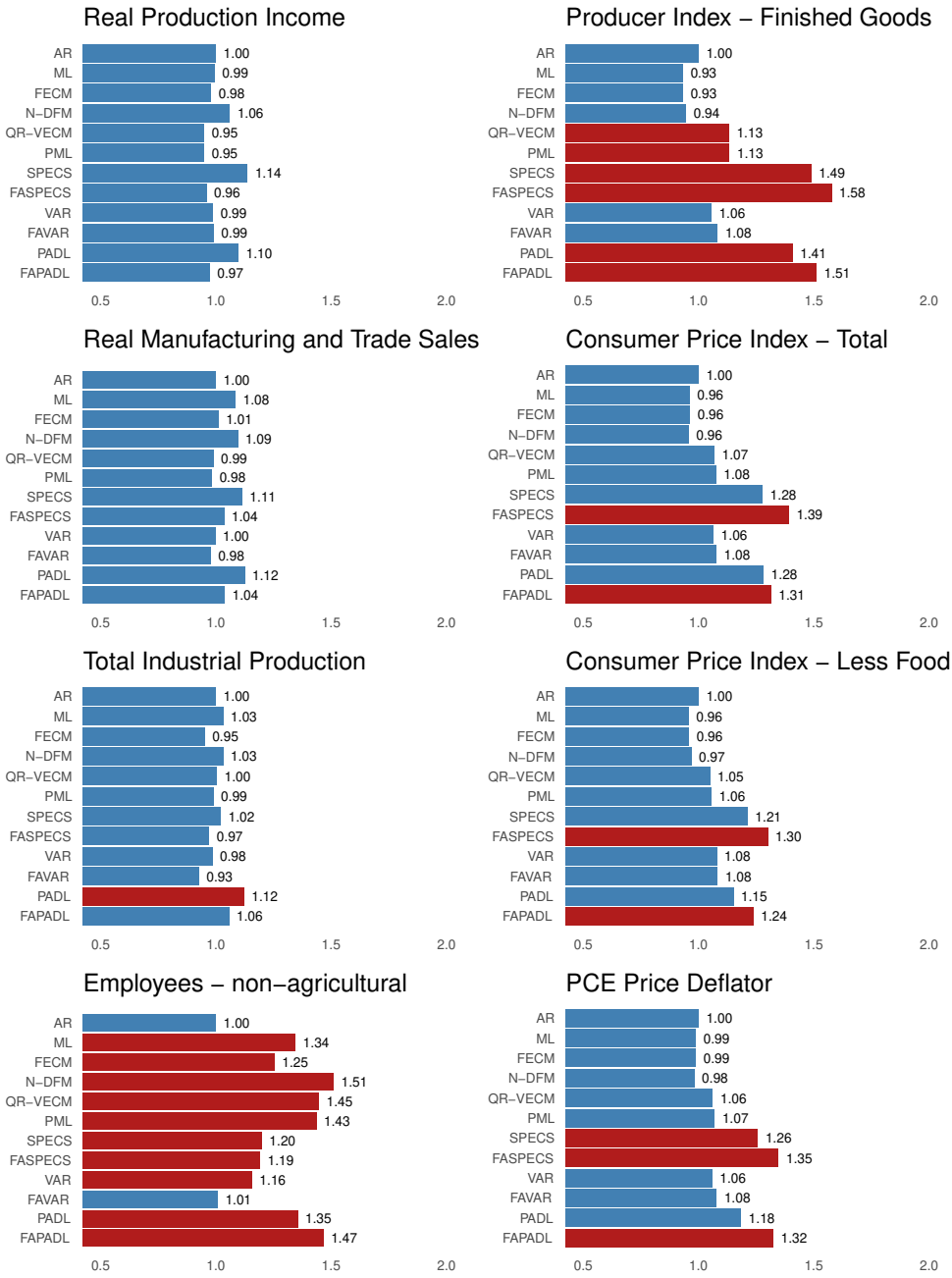

Figure 5.4: MCS and relative MSFEs for 1-month horizon. Methods that are included in the MCS are depicted as blue and methods that are excluded from the MCS are depicted in red.

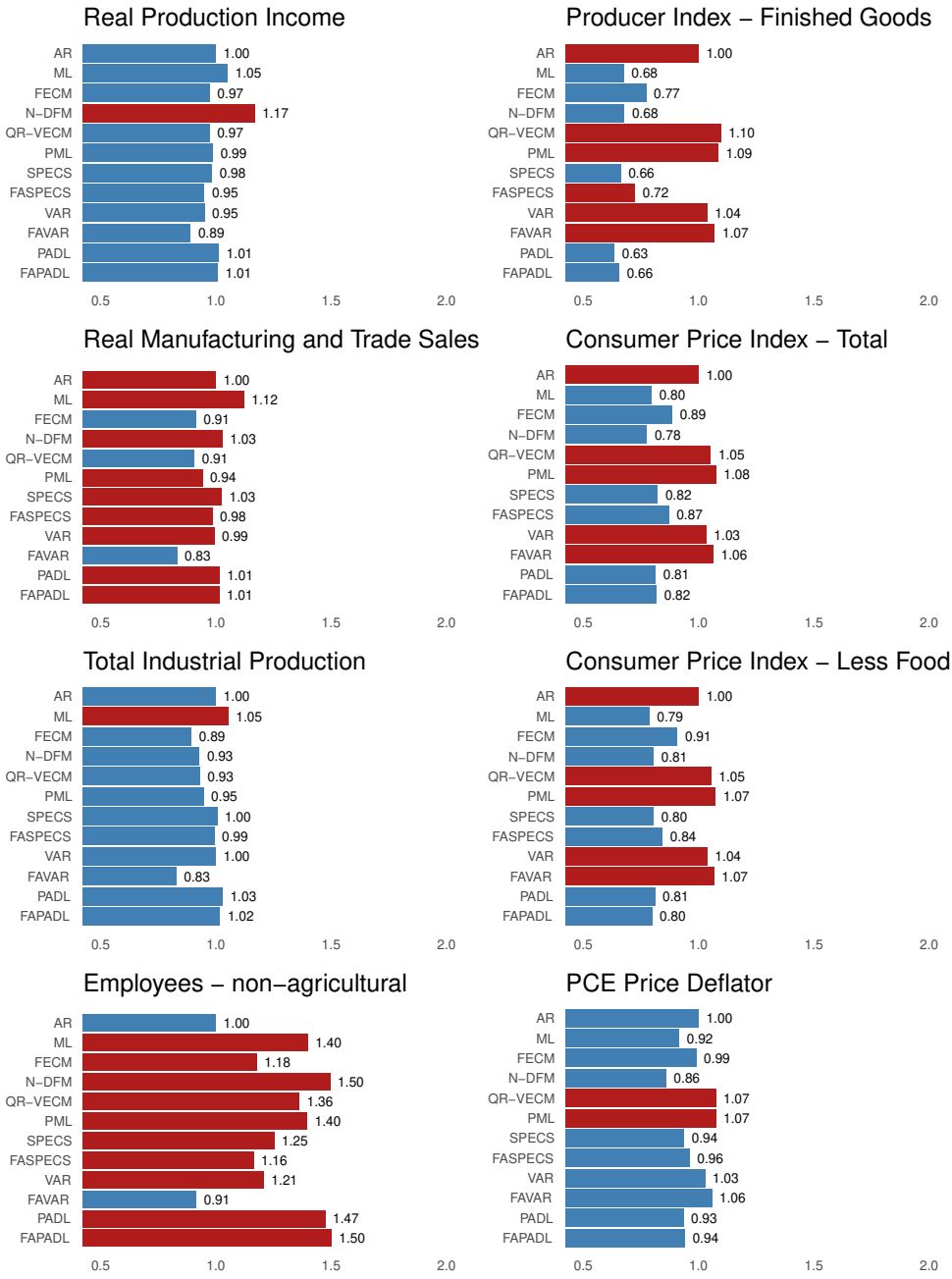

Figure 5.5: MCS and relative MSFEs for 6-month horizon. Methods that are included in the MCS are depicted as blue and methods that are excluded from the MCS are depicted in red.

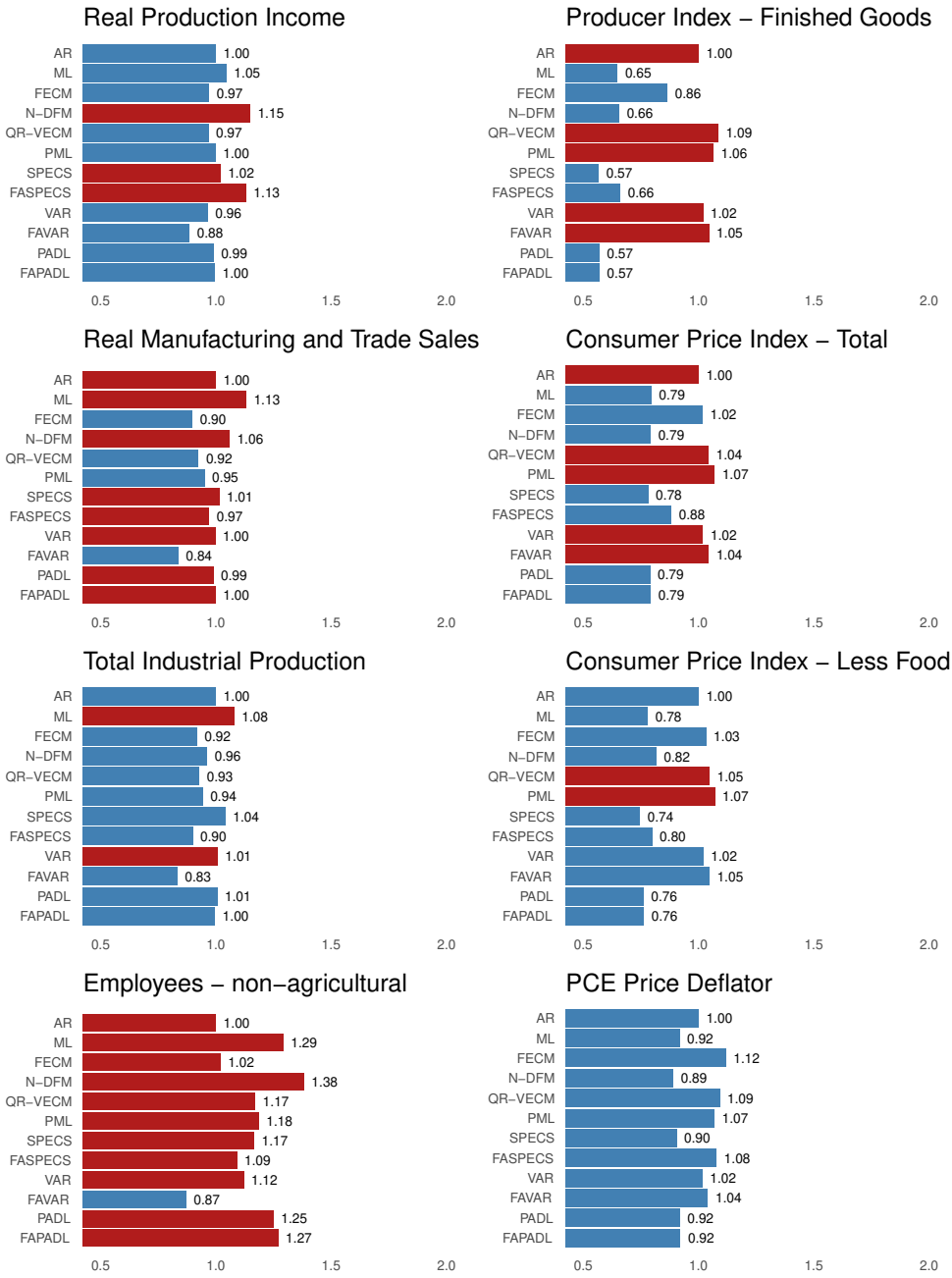

Figure 5.6: MCS and relative MSFEs for 12-month horizon. Methods that are included in the MCS are depicted as blue and methods that are excluded from the MCS are depicted in red.

Table 5.1 Overview of the variables included for QR-VECM and PML.

	FRED code	description
Real	RPI	Real Personal Income
	CMRMTSPLx	Real Manufacturing and Trades Industries Sale
	INDPRO	IP Index
	PAYEMS	All Employees: Total nonfarm
Nominal	WPSFD49207	PPI: Finished Goods
	CPIAUCSL	CPI : All Items
	CPIULFSL	CPI : All Items Less Food
	PCEPI	Personal Cons. Expend.: Chain Index
Additional	CUMFNS	Capacity Utilization: Manufacturing
	HWI	Help-Wanted Index for United States
	UNRATE	Civilian Unemployment Rate
	UEMPMEAN	Average Duration of Unemployment (Weeks)
	HOUST	Housing Starts: Total New Privately Owned
	PERMIT	New Private Housing Permits (SAAR)
	BUSINVx	Total Business Inventories
	M1SL	M1 Money Stock
	M2SL	M2 Money Stock
	FEDFUNDS	Effective Federal Funds Rate
	TB3MS	3-Month Treasury Bill
	GS5	5-Year Treasury Rate
	GS10	10-Year Treasury Rate
	EXJPUSx	Japan / U.S. Foreign Exchange Rate
	EXUSUKx	U.S. / U.K. Foreign Exchange Rate
	EXCAUSx	Canada / U.S. Foreign Exchange Rate
	S.P.500	S&P Common Stock Price Index: Composite

the four nominal series, the sparse high-dimensional methods display relatively poor performance, regardless of whether they take into account potential cointegration in the data. Overall, no clear distinction is visible between the non-stationary and stationary methods, although this may not come as a surprise given the short forecast horizon. As usual, the AR benchmark appears hard to beat and is not excluded from any of the model confidence sets here.

The forecast comparisons for longer forecast horizons display stronger differentiation across methods. Our findings are qualitatively similar for the 6-month and 12-month horizons, and, for the sake of brevity, we comment here on the 12-month horizon only. The results for the first three real series again do not portray a preference for taking into account cointegration versus transforming the data. Comparing VAR to FAVAR and ML to FECM, incorporating information across the whole dataset seems to positively affect forecast performance, a finding that is additionally

confirmed by the favourable performance of the penalized VECM estimators. The FAVAR substantially outperforms on the employment series, being the only method included in the model confidence set. On the nominal series, the single-equation methods perform well, again not showing any gain or loss in predictive power by accounting for cointegration. The ML and N-DFM procedure methods show favourable forecast accuracy as well, whereas the two penalized VECM estimators appear inferior on the nominal series. The AR benchmark is excluded for four out of eight series.

In summary, the comparative performance is strongly dependent on the choice of dependent variable and forecast horizon. For short forecast horizons, hardly any statistically significant differences in forecast accuracy are observed. However, for longer horizons the differences are more pronounced, with factor-augmented or penalized full system estimators performing well on the real series, the FAVAR strongly outperforming on the employment series, and the single-equation methods appearing superior on the nominal series. The findings do not provide conclusive evidence whether cointegration matters for forecasting.

5.5.2 Unemployment Nowcasting with Google Trends

In this section we revisit the nowcasting application of Chapter 3, where we consider nowcasting unemployment using Google Trends data. One of the advantages of modern high-dimensional datasets is that information obtained from internet activity is often available on very short notice, and can be used to supplement official statistics produced by statistical offices. For instance, internet searches about unemployment-related issues may contain information about people being or becoming unemployed, and could be used to obtain unemployment estimates before statistical offices are able to produce official unemployment statistics.

Google records weekly and monthly data on the popularity of specific search terms through its publicly available Google Trends service,[17] with data being available only days after a period ends. On the other hand, national statistical offices need weeks to process surveys and produce official unemployment figures for the preceding month. As such, Google Trends data on unemployment-related queries would appear to have the potential to produce timely nowcasts of the latest unemployment figures.

Indeed, Schiavoni et al. (2019) propose a dynamic factor model within a state space context to combine survey data with Google Trends data to produce more timely official umemployment statistics. They illustrate their method using a dataset

[17] `https://trends.google.com/trends`

of about one hundred unemployment-related queries in the Netherlands obtained from Google Trends. In Chapter 3, we consider a similar setup with the same Google Trends data, but consider the conceptually simpler setup where the dependent variable to be nowcasted is the official published unemployment by Statistics Netherlands.[18] Moreover, they exclusively focus on penalized regression methods. In this section we revisit their application in the context of the methods discussed in this chapter. For full details on the dataset, which is available on the website of the author of this thesis, we refer to Chapter 3.

Transformations to stationarity

As for the FRED-MD dataset, we first consider the different ways to classify the series into I(0), I(1) and I(2) series. However, unlike for the FRED data, here we don't have a pre-set classification available, and therefore unit root testing is a necessity before continuing the analysis. Moreover, as the dataset could easily be extended to an arbitrarily high dimension by simply adding other relevant queries, an automated fully-data driven method is required.

This lack of a known classification also means that our first strategy as used in Section 5.5.1 has to be adapted, as we cannot differ I(2) series a priori. In particular, for our first strategy we assume that the series can be at most I(1), and hence we perform only a single unit root test on the levels of all series. Our second strategy is again the Pantula principle as in Section 5.5.1. Within each strategy we consider the same four tests as before.[19]

The classification results are given in Figure 5.7. Generally they provide strong evidence that nearly all series are I(1), with most methods only finding very few I(0) and I(2) series. Interestingly, one of the few series that the methods disagree about is the unemployment series, which receives all three possible classifications. From our previous results we may expect this series, our dependent variable, to be the major determinant of forecast accuracy. Aside from this result, the most striking result is the performance of the naive tests, that find many more I(0) variables than the other methods. One possible explanation for this result may be the nature of the Google Trends data, that can exhibit large changes in volatility. As standard unit root tests are not robust to such changes, a naive strategy might seriously be affected, as appears to be the case here.

[18] Additionally, this means the application does not require the use of the private survey data and is based on publicly available data only.

[19] As Google reports the search frequencies in relative terms (both to the past and other searches), we do not take logs anywhere.

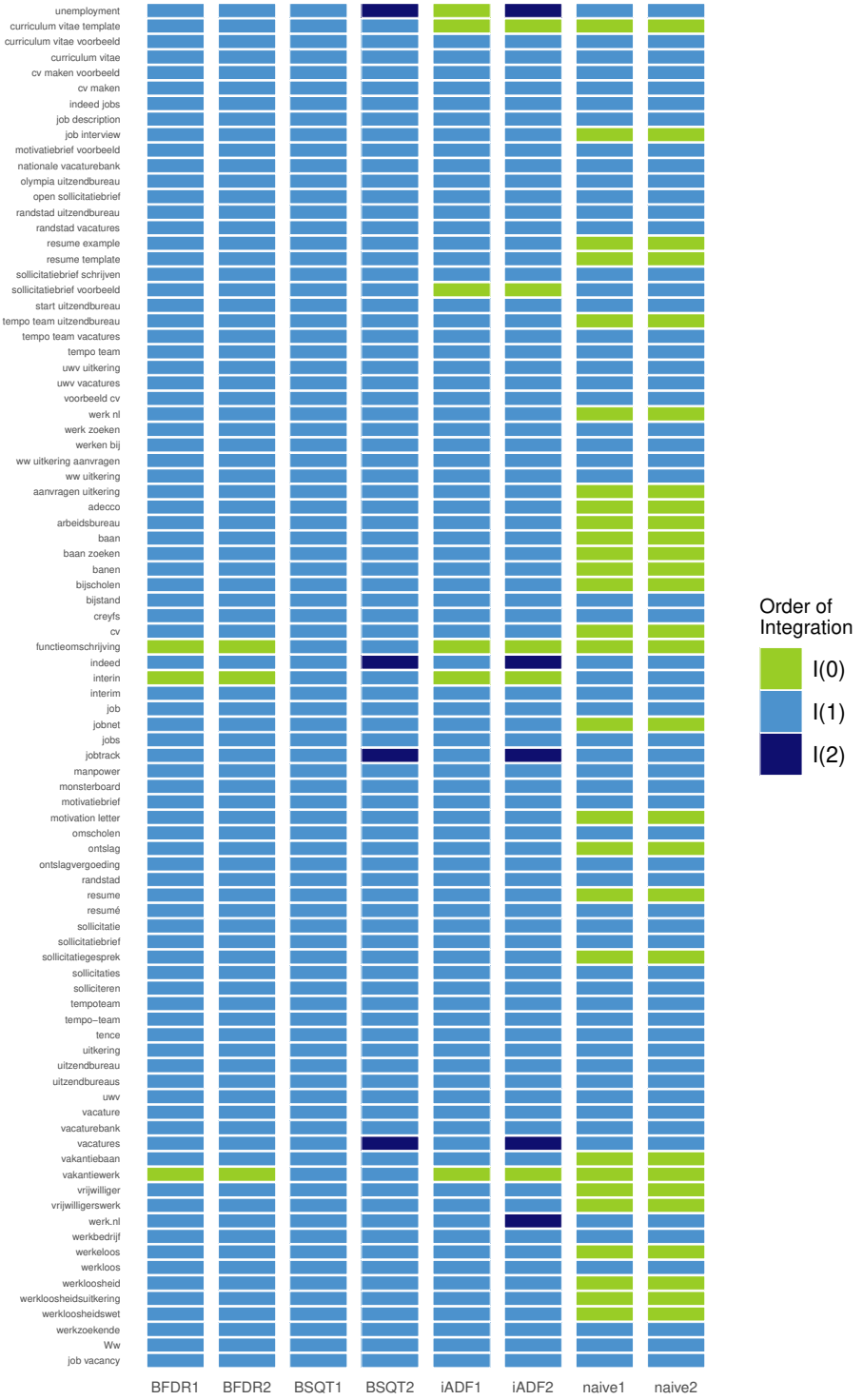

Figure 5.7: Classification of integration order of unemployment dataset.

Forecast comparison

We now compare the nowcasting performance of the high-dimensional methods. Given our focus on forecasting the present, that is $h = 0$, for a single variable, there is little benefit in considering the system estimators we used before. Therefore we only consider the subset of single-equation models that allow for nowcasting. Specifically, we include SPECS as described in Section 5.4.2 as well as its modification FA-SPECS described in Section 5.5.1 as methods that explicitly account for unit roots and cointegration. Furthermore, we include PADL and FAPADL as described in Section 5.5.1. For all methods, the modification for nowcasting is done by setting $h = 0$, where we implicitly assume that at time t the values for the explanatory variables are available, but that for unemployment is not. This corresponds to the real-life situation.

For SPECS we model unemployment as (at most) $I(1)$, given that this is its predominant classification in Figure 5.7. Additionally, we include all regressors in levels, thereby implicitly assuming these are at most $I(1)$ as well, which is again justified by the preceding unit root tests. For PADL and FAPADL we transform the series to stationarity according to the obtained classifications. Again we consider an AR model as benchmark, while all other implementational details are the same as in Section 5.5.1.

Our dataset covers monthly data from January 2004 until December 2017 for unemployment obtained from Statistics Netherlands, and 87 Google Trends series. We estimate the models on a rolling window of 100 observations each, leaving 64 time periods for obtaining nowcasts. We compare the nowcast accuracy through relative Mean Squared Nowcast Error (MSNE), with the AR model as benchmark, and obtain 90% Model Confidence Sets containing the best models in the same way as in Section 5.5.1.

Figure 5.8 presents the results. We see that, with the exception of the PADL - iADF1 method, all methods outperform the AR benchmark, although the 90% MCS does not find the differences to be significant. Factor augmentation generally leads to slightly more accurate forecasts than the full penalization approaches, but differences are marginal. Interestingly, the classification of unemployment appears to only have a minor effect on the accuracy, with $I(0)$, $I(1)$ and $I(2)$ classifications all performing similarly. This does not necessarily contradict the results in Section 5.5.1, as differences were only pronounced there for longer forecast horizons, whereas the forecast horizon here is immediate. Finally, we observe that the SPECS methods are always at least as accurate as their counterparts that do not take cointegration

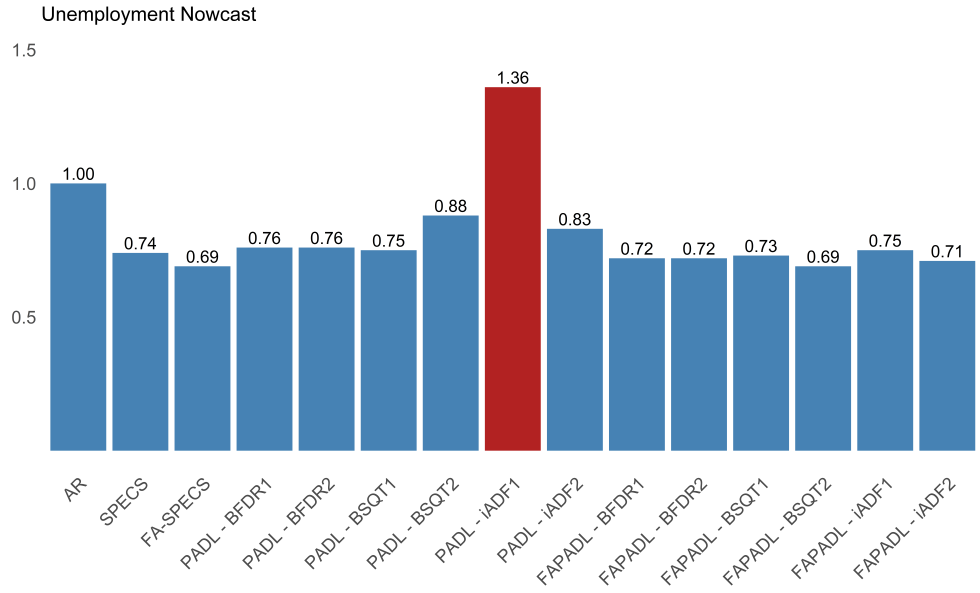

Figure 5.8: MCS and relative MSNEs for the unemployment nowcasts. Methods that are included in the MCS are depicted as blue and methods that are excluded from the MCS are depicted in red.

into account. It therefore seems to pay off to allow for cointegration, even though differences are again marginal.

5.6 Conclusion

In this chapter we investigated how the potential presence of unit roots and cointegration impacts macroeconomic forecasting in the high-dimensional setting. We considered both the strategies of transforming all data to stationarity, and of explicitly modelling any unit roots and cointegrating relationships.

The strategy of transforming to stationarity is commonly thought of as allowing one to bypass the unit root issue. However, this strategy is not innocuous as often thought, as it still relies on a correct classification of the orders of integration of all series. Given that this needs to be done for a large number of series, there is a lot of room for errors, and naive unit root testing is not advised. We discussed potential pitfalls for this classification, and evaluated methods designed to deal with issues of poor size and power of unit root tests, as well as controlling appropriate error rates in multiple testing.

Next we considered modelling unit roots and cointegration directly in a high-dimensional framework. We reviewed methods approaching the problem from two different philosophies, namely that of factor models and that of penalized regression. Within these philosophies we also highlighted differences among the proposed methods both in terms of underlying assumptions and implementation issues.

We illustrated these methods in two empirical applications; the first considered forecasting macroeconomic variables using the well-established FRED-MD dataset, while the second considered nowcasting unemployment using Google Trends data. Both applications showed that transforming to stationarity requires careful considerations of the methods used. While the specific method used for accounting for multiple testing generally only led to marginal differences, a correct classification of the variable to be forecasted is critically important. We therefore recommend paying specific attention to these variables by, for example, performing the classification using multiple approaches to ensure that the classification found is credible.

The applications also demonstrated that there is no general way to model cointegration that is clearly superior. Indeed, the results do not show a clear conclusion on whether cointegration should be taken into account. This result, perhaps unsurprisingly, mirrors the literature on low-dimensional time series. It therefore remains up to the practitioner to decide for their specific application if, and if yes how, cointegration should be modelled for forecasting purposes. Overall, the methods we consider in this chapter provide reliable tools to do so, should the practitioner wish to do so.

Concluding, several reliable tools are available for dealing with unit roots and cointegration in a high-dimensional forecasting setting. However, there is no panacea; a single best approach that is applicable in all settings does not exist. Instead, dealing with unit roots and cointegration in practice requires careful consideration and investigation which methods are most applicable in a given particular application. We also note that the field is rapidly developing, and major innovations are still to be expected in the near future. For instance, interval or density forecasting in high-dimensional systems with unit roots remains an entirely open issue. As high-dimensional inference is already complicated by issues such as post-selection bias, extending this to the unit root setting is very challenging indeed. Such tools however will be indispensable for the macroeconomic practitioner, and therefore constitute an exciting avenue for future research.

Chapter 6

Conclusion

"If you want to assert a truth, first make sure it is not just an opinion that you desperately want to be true."

- Neil deGrasse Tyson (1958 - present)

This chapter concludes the thesis as a whole. We first provide some general conclusions that can be drawn from this work. As each chapter is annotated with its own conclusion, we provide a holistic overview here and refer the reader to the individual chapters for details. The chapter, and therewith the thesis, ends with a discussion of some limitations and prospective avenues for future research.

The main result brought forward in this thesis, is that penalized regression offers substantial theoretical and empirical advantages in high-dimensional (non-)stationary time series settings. Throughout the thesis, it has been demonstrated that penalized regression techniques offer competitive predictive performance relative to a wide variety of factor models, which have long constituted the pre-dominant modelling strategy on large (macro)economic datasets. However, naive application of penalized regression to non-stationary datasets in levels is not insensitive to the well-known issue of spurious regression. We show that this problem can be circumvented by choosing an appropriate model specification that automatically takes into account the (co)integration properties of the data. Certain important choices, such as whether to correct for unit roots or to model cointegration directly, as well as whether a factor model or penalized regression method is most appropriate, remain application-dependent. Notwithstanding, our results demonstrate that lasso-type estimation may now be considered as a standard tool with wide applicability by the time series econometrician.

In Chapter 2, we show that penalized regression offers competitive predictive performance relative to factor models on stationary datasets. By means of simulations, we show in a controlled environment that lasso-type estimators provide superior forecast accuracy on sparse DGPs. Unsurprisingly, when the true DGP contains a 'well-behaved' factor structure, factor models arise as the superior modelling strategy, although the performance gain over penalized regression is marginal. More interestingly, when the DGP indeed possesses a factor structure, but with strong cross-sectional correlation in the idiosyncratic component, we observe that (i) the factors are estimated with poor accuracy, (ii) factor selection criteria fail to choose the correct number of factors, and (iii) the predictive performance of factor models turns out inferior to that of penalized regression. An empirical forecasting application to the famous FRED-MD dataset shows mixed results, with neither modelling philosophy consistently outperforming one another. Finally, our simulation results in the non-stationary setting highlight that lasso-type estimation on the dataset in levels is not insensitive to spurious regression, since a large number of irrelevant integrated variables are frequently included in the estimated model.

In recognition of the documented risk of spurious regression, Chapter 3 develops the single-equation penalized error correction selector (SPECS) as an automated estimation procedure for modelling (co)integrated datasets. SPECS is based on a single-equation model that, contrary to Chapter 2, is obtained from a VECM specification, rather than a stationary VAR. Consequently, variables that are stationary or integrated of order one are both admissible in the model. Our theoretical results, derived in a fixed-dimensional setting, show that SPECS possesses the oracle property. Furthermore, elaborate simulation results and an empirical application to nowcasting Dutch unemployment based on Google trends provide additional evidence of the favourable performance of penalized regression in non-stationary settings.

With the aim of providing better asymptotic approximations for high-dimensional applications, Chapter 4 extends the theoretical results of Chapter 3 to a framework in which the number of variables diverges along with the sample size. We show that SPECS maintains selection and estimation consistency in a high-dimensional setting and describe the inverse relationship between the rate at which the dimension diverges and the convergence rate of the estimator.

In recognition of the availability of high-dimensional estimators for both stationary and non-stationary datasets, Chapter 5 examines the issue of unit root testing in high dimensions and the performance differentials one may expect in both worlds. In general, the specific multiple hypothesis testing strategy by which to identify unit

roots only marginally affects the forecast performance of the methods. However, classification of the order of integration of the dependent variable has a substantial impact on the predictive performance of all methods included in the comparison, and therefore merits extra careful consideration by the researcher. The predictive performance comparisons on two empirical applications are not conclusive on whether cointegration should be taken into account for forecasting. Moreover, no estimation method consistently arise as superior. Perhaps unsurprisingly, careful consideration as to which modelling strategy is most appropriate for a particular application remains a necessity.

While this thesis offers new insights into the potential of penalized regression in high-dimensional time series analysis, it is far from complete. Accordingly, we proceed by suggesting several directions in which to results may be extended.

First, our theoretical results are based on pointwise convergence, thereby preventing uniformly valid inference. It is now well-recognized that post-model selection inference is complicated by the issue of post-selection bias and numerous solutions, such as post-double selection (Belloni et al., 2014) or the desparsified lasso (Van de Geer et al., 2014), have been proposed. However, none of these approaches extend easily to general stationary time series setttings, and extensions to the unit root setting are expected to be highly complicated. Nonetheless, the theoretical results of Chapter 3-4 may prove useful as intermediary results in the pursuit of uniformly valid post-model selection inference.

Second, the generality of the high-dimensional asymptotic framework in the non-stationary setting is hampered by the absence of knowledge regarding the behaviour of the minimum eigenvalues of sample covariance matrices based on integrated variables. Extending concepts such as the compatibility condition to the non-stationary setting, may allow for the lasso to be theoretically justifiable in higher-dimensional settings than the one proposed in Chapter 4.

Third, the shrinkage estimators in this thesis are largely limited to lasso-type estimators. While the (adaptive) lasso has arguably developed into the most popular form of penalized regression for variable selection, recent literature has proposed several prospective extensions or alternative estimators that are not considered in this thesis. For example, Belloni et al. (2011) and Belloni et al. (2014) propose the square-root lasso, which attains near-optimal oracle rates under less stringent assumptions than the plain lasso. Perhaps more importantly, the square-root lasso is 'self-tuning', thereby removing the burden of manually selecting the desired degree of penalization.

Their results rely on conditions on the Gram matrix that are similar in spirit to the compatibility condition, thereby further illustrating the importance of deriving such conditions for the non-stationary setting. Alternatively, Fan and Li (2001) approach the topic of variable selection more generally and argue that the lasso can be seen as part of a larger class of estimators referred to as non-concave penalized maximum likelihood estimators. While some penalties result in non-convex estimation procedures for which multiple local solutions exist, Fan et al. (2014) propose a procedure called folded non-concave penalization which provides an approximation scheme that is able to recover the oracle solutions under mild conditions. Since this procedure does not rely on initial estimates, these methods provide yet another interesting direction in which to generalize our results.

Fourth, extending the variable selection properties of the lasso to selection of the deterministic specification allows for even greater automation in the model building process. Chapter 3 takes into account potential deterministic variables in the model by de-meaning and de-trending. While this approach results in a limit distribution that is insensitive to the presence of a non-zero constant or deterministic trend, it does not necessarily result in the most efficient estimator. Selection of the trend component, however, is a non-trivial extension, because a deterministic trend dominates the stochastic variation and leads to asymptotically singular covariance matrices. Nonetheless, automated selection of the deterministic specification would clearly benefit the applied researcher.

Finally, we have mainly focussed on a specific form of non-stationarity, namely variables that are integrated of order one. However, it remains a debated issue whether certain macroeconomic variables, such as price indices, are integrated of order two (see for example the discussion in Marcellino et al., 2006a). Therefore, a careful consideration of the effects of (misspecification of) variables that are integrated of higher orders, as well as model extensions that accommodate such variables, would be a valuable contribution to the literature.

Concluding, while shrinkage estimation has already shown great potential in time series settings, there are still a lot of outstanding issues that could be tackled by developing new methods, or increasing our understanding of existing methods. We hope that the results brought forward in this thesis contribute to our general understanding of penalized regression, and prove useful for the applied researcher in modelling large time series datasets, as well as the theoretical researcher in further developing the theory in relevant time series setting.

Bibliography

Abadir, K. and J. Magnus (2002). Notation in econometrics: a proposal for a standard. *The Econometrics Journal 5*, 76–90.

Abadir, K. M. and J. R. Magnus (2005). *Matrix algebra*, Volume 1. Cambridge University Press.

Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*(3), 1203–1227.

Ahn, S. K. and G. C. Reinsel (1990). Estimation for partially nonstationary multivariate autoregressive models. *Journal of the American Statistical Association 85*(411), 813–823.

Akesson, F. and J. Lehoczky (1998). Discrete eigenfunction expansion of multidimensional Brownian motion and the Ornstein-Uhlenbeck process. Preprint.

Alessi, L., M. Barigozzi, and M. Capasso (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics & Probability Letters 80*(23), 1806–1813.

Artis, M. J., A. Banerjee, and M. Marcellino (2005). Factor forecasts for the UK. *Journal of Forecasting 24*, 279–298.

Bai, J. (2004). Estimating cross-section common stochastic trends in nonstationary panel data. *Journal of Econometrics 122*(1), 137–183.

Bai, J., K. Li, and L. Lu (2016). Estimation and inference of FAVAR models. *Journal of Business & Economic Statistics 34*, 620–641.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*, 191–221.

Bai, J. and S. Ng (2004). A PANIC attack on unit roots and cointegration. *Econometrica 72*(4), 1127–1177.

Bai, J. and S. Ng (2008a). Forecasting economic time series using targeted predictors. *Journal of Econometrics 146*, 304–217.

Bai, J. and S. Ng (2008b). Large dimensional factor analysis. *Foundations and Trends in Econometrics 3*, 89–163.

Bai, Z.-D. and Y.-Q. Yin (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Annals of Probability 21*, 1275–1294.

Bańbura, M., D. Giannone, and L. Reichlin (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics 25*, 71–92.

Banerjee, A., J. Dolado, and R. Mestre (1998). Error-correction mechanism tests for cointegration in a single-equation framework. *Journal of Time Series Analysis 19*(3), 267–283.

Banerjee, A. and Marcellino (2009). Factor-augmented error correction models. In J. L. Castle and N. Shephard (Eds.), *The Methodology and Practice of Econometrics - A Festschrift for David Hendry*, pp. 589–612. Oxford: Oxford University Press.

Banerjee, A., M. Marcellino, and I. Masten (2014a). Forecasting with factor-augmented error correction models. *International Journal of Forecasting 30*(3), 589–612.

Banerjee, A., M. Marcellino, and I. Masten (2014b). Forecasting with factor-augmented error correction models. *International Journal of Forecasting 30*(3), 589–612.

Banerjee, A., M. Marcellino, and I. Masten (2016). An overview of the factor-augmented error-correction model. In E. Hillebrand and S. J. Koopman (Eds.), *Dynamic Factor Models*, Volume 35 of *Advances in Econometrics*, Chapter 1, pp. 3–41. Emerald Group Publishing Limited.

Banerjee, A., M. Marcellino, and I. Masten (2017). Structural FECM: Cointegration in large-scale structural FAVAR models. *Journal of Applied Econometrics 32*(6), 1069–1086.

Barigozzi, M. and C. Brownlees (2019). NETS: Network estimation for time series. *Journal of Applied Econometrics 34*(3), 347–364.

Barigozzi, M., M. Lippi, and M. Luciani (2016a). Dynamic factor models, cointegration, and error correction mechanisms. arXiv e-print 1510.02399.

Barigozzi, M., M. Lippi, and M. Luciani (2016b). Non-stationary dynamic factor models for large datasets. Working Paper.

Barigozzi, M., M. Lippi, and M. Luciani (2017). Dynamic factor models, cointegration, and error correction mechanisms. arXiv e-prints 1510.02399, arXive.

Barigozzi, M., M. Lippi, and M. Luciani (2018). Non-stationary dynamic factor models for large datasets. arXiv e-prints 1602.02398, arXive.

Barigozzi, M. and L. Trapani (2018). Determining the dimension of factor structures in non-stationary large datasets. arXiv e-prints 1806.03647, arXive.

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*(2), 521–547.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*, 608–650.

Belloni, A., V. Chernozhukov, and L. Wang (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika 98*, 791–806.

Belloni, A., V. Chernozhukov, L. Wang, et al. (2014). Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics 42*(2), 757–788.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B 57*(1), 289–300.

Bergmeir, C., R. J. Hyndman, B. Koo, et al. (2015). A note on the validity of cross-validation for evaluating time series prediction. *Monash University Department of Econometrics and Business Statistics Working Paper 10*, 15.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013). Valid post-selection inference. *Annals of Statistics 41*, 802–837.

Bernanke, B., J. Boivin, and P. Eliasz (2005a). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *Quarterly Journal of Economics 120*(1), 387–422.

Bernanke, B. S., J. Boivin, and P. Eliasz (2005b). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics 120*(1), 387–422.

Bernardini, E. and G. Cubadda (2015). Macroeconomic forecasting and structural analysis through regularized reduced-rank regression. *International Journal of Forecasting 31*(3), 682–691.

Boivin, J. and S. Ng (2006). Are more data always better for factor analysis? *Journal of Econometrics 132*, 169–194.

Boswijk, H. P. (1994). Testing for an unstable root in conditional and structural error correction models. *Journal of Econometrics 63*, 37–60.

Brockwell, P. J. and R. A. Davis (1991). *Time Series: Theory and Methods* (2nd ed.). New York: Springer-Verlag.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.

Buono, D., G. L. Mazzi, G. Kapetanios, M. Marcellino, and F. Papailias (2017). Big data types for macroeconomic nowcasting. *Eurostat Review on National Accounts and Macroeconomic Indicators 1*(2017), 93–145.

Callot, L. A. and A. B. Kock (2014). Oracle efficient estimation and forecasting with the adaptive lasso and the adaptive group lasso in vector autoregressions. In N. Haldrup, M. Meitz, and P. Saikkonen (Eds.), *Essays in Nonlinear Time Series Econometrics*, Chapter 10, pp. 238–268. Oxford: Oxford University Press.

Campbell, J. Y. and R. J. Shiller (1987). Cointegration and tests of present value models. *Journal of political economy 95*(5), 1062–1088.

Casella, G. and R. L. Berger (2002). *Statistical Inference*, Volume 2. Duxbury Pacific Grove, CA.

Cavaliere, G. (2005). Unit root tests under time-varying variances. *Econometric Reviews 23*(3), 259–292.

Cavaliere, G., P. C. B. Phillips, S. Smeekes, and A. M. R. Taylor (2015). Lag length selection for unit root tests in the presence of nonstationary volatility. *Econometric Reviews 34*(4), 512–536.

Cavaliere, G. and A. M. R. Taylor (2008). Bootstrap unit root tests for time series with nonstationary volatility. *Econometric Theory 24*(1), 43–71.

Cavaliere, G. and A. M. R. Taylor (2009). Bootstrap M unit root tests. *Econometric Reviews 28*(5), 393–421.

Chamberlain, G. and M. Rothschild (1983). Factor structure, and mean-variance analysis on large asset markets. *Econometrica 51*, 1281–1304.

Chen, X., M. Xu, and W. B. Wu (2013). Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics 41*, 2994–3021.

Cheng, X. and P. C. B. Phillips (2009). Semiparametric cointegrating rank selection. *Econometrics Journal 12*(suppl1), S83–S104.

Chernozhukov, V., W. K. Härdle, C. Huang, and W. Wang (2018). LASSO-driven inference in time and space. arXiv e-print 1806.05081, arXive.

Choi, H. and H. Varian (2012). Predicting the present with Google Trends. *Economic Record 88*(s1), 2–9.

Choi, I. (2015). *Almost All About Unit Roots: Foundations, Developments, and Applications*. Cambridge University Press.

Chortareas, G. and G. Kapetanios (2009). Getting PPP right: Identifying mean-reverting real exchange rates in panels. *Journal of Banking and Finance 33*(2), 390–404.

Chou, W., K. F. Denis, and C. F. Lee (1996). Hedging with the nikkei index futures: The convential model versus the error correction model. *The Quarterly Review of Economics and Finance 36*(4), 495–505.

Christoffersen, P. F. and F. X. Diebold (1998). Cointegration and long-horizon forecasting. *Journal of Business & Economic Statistics 16*(4), 450–456.

Clements, M. P. and D. F. Hendry (1995). Forecasting in cointegrated systems. *Journal of Applied Econometrics 10*(2), 127–146.

Croux, C. and P. Exterkate (2011). Sparse and robust factor modelling. Technical report, Tinbergen Institute Discussion Paper TI 122/4.

Davidson, J. (1994). *Stochastic limit theory: An introduction for econometricians*. Oxford University Press.

Davidson, J. (2000). *Econometric Theory* (2nd ed.). Oxford: Blackwell Publishers.

De Mol, C., D. Giannone, and L. Reichlin (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics 146*, 318–328.

Dickey, D. A. and W. A. Fuller (1979). Distribution of estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association 74*(366a), 427–431.

Diebold, F. X. (2012). On the origin(s) and development of the term 'Big Data'. Working paper, PIER.

Diebold, F. X. and L. Kilian (2000). Unit-root tests are useful for selecting forecasting models. *Journal of Business & Economic Statistics 18*(3), 265–273.

Doz, C., D. Giannone, and L. Reichlin (2011). A two-step estimator for large approximate dynamic factor models based on kalman filtering. *Journal of Econometrics 164*(1), 188–205.

Doz, C., D. Giannone, and L. Reichlin (2012). A quasi–maximum likelihood approach for large, approximate dynamic factor models. *Review of Economics and Statistics 94*, 1014–1024.

Dwyer Jr, G. P. and M. S. Wallace (1992). Cointegration and market efficiency. *Journal of International Money and Finance 11*(4), 318–327.

Eickmeier, S. and C. Ziegler (2008). How successful are dynamic factor models at forecasting output and inflation? A meta-analytic approach. *Journal of Forecasting 27*, 237–265.

Elliott, G., T. J. Rothenberg, and J. H. Stock (1996). Efficient tests for an autoregressive unit root. *Econometrica 64*(4), 813–836.

Enders, W. (2008). *Applied econometric time series* (4th ed.). John Wiley & Sons.

Engle, R. F. and C. W. J. Granger (1987). Co-integration and error correction: representation, estimation and testing. *Econometrica: Journal of the Econometric Society 55*, 251–276.

Engle, R. F. and B. S. Yoo (1987). Forecasting and testing in co-integrated systems. *Journal of Econometrics 35*(1), 143–159.

Fan, J. and R. Li (2001). Variable selection via nonconcace penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J., L. Xue, and H. Zou (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of statistics 42*, 819.

Forni, M., D. Giannone, M. Lippi, and L. Reichlin (2009). Opening the black box: Structural factor models with large cross sections. *Econometric Theory 25*(05), 1319–1347.

Forni, M., A. Giovannelli, M. Lippi, and S. Soccorsi (2018). Dynamic factor model with infinite-dimensional factor space: Forecasting. *Journal of Applied Econometrics 33*, 625–642.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized dynamic factor model: identification and estimation. *The Review of Economics and Statistics 82*, 540–554.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2005a). The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association 100*(471), 830–840.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2005b). The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association 100*(471), 830–840.

Forni, M., M. Hallin, M. Lippi, and P. Zaffaroni (2015). Dynamic factor models with infinite-dimensional factor spaces: one-sided representations. *Journal of Econometrics 185*(2), 359–371.

Franses, P. H. and M. McAleer (1998). Testing for unit roots and non-linear transformations. *Journal of Time Series Analysis 19*(2), 147–164.

Friedman, J. H., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 1–22.

Friedrich, M., S. Smeekes, and J.-P. Urbain (2018). Autoregressive wild bootstrap inference for nonparametric trends. arXiv e-prints 1807.02357, arXive.

Garcia, M. G., M. C. Medeiros, and G. F. Vasconcelos (2017). Real-time inflation forecasting with high-dimensional models: The case of Brazil. *International Journal of Forecasting 33*, 679–693.

Gelper, S. and C. Croux (2008). Least angle regression for time series forecasting with many predictors. Working paper, KU Leuven-Faculty of Business and Economics.

Giannone, D., L. Reichlin, and D. Small (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics 55*(4), 665–676.

Gonçalves, S. and B. Perron (2014). Bootstrapping factor-augmented regression models. *Journal of Econometrics 182*(1), 156–173.

Hallin, M. and R. Liška (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association 102*(478), 603–617.

Hamilton, J. D. (1994). *Time series analysis* (1st ed.). Princeton University Press.

Hanck, C. (2009). For which countries did PPP hold? A multiple testing approach. *Empirical Economics 37*(1), 93–103.

Hansen, C. and Y. Liao (2019). The factor-lasso and k-step bootstrap approach for inference in high-dimensional economic applications. *Econometric Theory 35*, 465–509.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

Harvey, D. I., S. J. Leybourne, and A. M. R. Taylor (2009). Unit root testing in practice: dealing with uncertainty over the trend and initial condition. *Econometric Theory 25*(3), 587–636.

Harvey, D. I., S. J. Leybourne, and A. M. R. Taylor (2012). Testing for unit roots in the presence of uncertainty over both the trend and initial condition. *Journal of Econometrics 169*(2), 188–195.

Hastie, T., R. Tibshirani, and J. Friedman (2008). *The Elements of Statistical Learning*. Springer.

Hastie, T., M. Wainwright, and R. Tibshirani (2015). *Statistical Learning with Sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Hendry, D. F. and K. Juselius (2001). Explaining cointegration analysis: Part ii. *The Energy Journal 22*(1).

Horn, R. A., C. R. Johnson, and L. Elsner (1994). *Topics in Matrix Analysis*. Cambridge University Press.

Hsu, N., H. Hung, and Y. Chang (2008). Subset selection for vector autoregressive processes using lasso. *Computational Statistics and Data Analysis 52*, 3645–3657.

Huang, J., S. Ma, and C.-H. Zhang (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica 18*, 1603–1618.

Hyndman, R. J. (2016). *forecast: Forecasting functions for time series and linear models*. R package. R package version 7.2.

Hyndman, R. J. and G. Athanasopoulos (2018). *Forecasting: principles and practice*. OTexts.

Jiang, W. (2009). On uniform deviations of general empirical risks with unboundedness, dependence, and high dimensionality. *Journal of Machine Learning Research 10*, 977–996.

Johansen, S. (1992a). Cointegration in partial systems and the efficiency of single-equation analysis. *Journal of Econometrics 52*(3), 389–402.

Johansen, S. (1992b). Testing weak exogeneity and the order of cointegration in uk money demand data. *Journal of Policy modeling 14*(3), 313–334.

Johansen, S. (1995a). *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford University Press.

Johansen, S. (1995b). A stastistical analysis of cointegration for i(2) variables. *Econometric Theory 11*(1), 25–59.

Johansen, S. and A. R. Swensen (2004). More on testing exact rational expectations in cointegrated vector autoregressive models: Restricted constant and linear term. *The Econometrics Journal 7*(2), 389–397.

Jolliffe, I. T., N. T. Trendafilov, and M. Uddin (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics 12*(3), 531–547.

Juselius, K. and R. MacDonald (2004). International parity relationships between the usa and japan. *Japan and the World economy 16*(1), 17–34.

Justiniano, A. and G. Primiceri (2008). The time-varying volatility of macroeconomic fluctuations. *American Economic Review 98*(3), 604–641.

Kascha, C. and C. Trenkler (2015). Forecasting VARs, model selection and shrinkage. Working paper 15-07, University of Mannheim / Department of Economics.

Kim, H. H. and N. R. Swanson (2014). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics 178*, 352–367.

Klaassen, S., J. Kueck, and M. Spindler (2017). Transformation models in high-dimensions. arXiv e-prints 1712.07364, arXive.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics 28*, 1356–1378.

Kock, A. B. (2016). Consistent and conservative model selection with the adaptive lasso in stationary and nonstationary autoregressions. *Econometric Theory 32*, 243–259.

Kock, A. B. and L. Callot (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics 186*, 325–344.

Kramer, W. and L. Davies (2002). Testing for unit roots in the context of misspecified logarithmic random walks. *Economics Letters 74*(3), 313–319.

Kristensen, J. T. (2017). Diffusion indexes with sparse loadings. *Journal of Business & Economic Statistics 35*, 434–451.

Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014). The parable of Google Flu: traps in big data analysis. *Science 343*, 1203–1205.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics 44*, 907–927.

Lee, J. H., Z. Shi, and Z. Gao (2018). On LASSO for predictive regressio. arXiv e-prints 1810.03140, arXive.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*(1), 21–59.

Li, J. and W. Chen (2014). Forecasting macroeconomic time series: Lasso-based approaches and their forecast combinations with dynamic factor models. *International Journal of Forecasting 30*, 995–1015.

Liang, C. and M. Schienle (2019). Determination of vector error correction models in high dimensions. *Journal of Econometrics 208*(2), 418–441.

Liao, Z. and P. C. B. Phillips (2015). Automated estimation of vector error correction models. *Econometric Theory 31*, 581–646.

Luciani, M. (2014). Forecasting with approximate dynamic factor models: The role of non-pervasive shocks. *International Journal of Forecasting 30*(1), 20–29.

Ludvigson, S. C. and S. Ng (2009). A factor analysis of bond risk premia. Nber working paper no. 15188, National Bureau of Economic Research.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media.

Marcellino, M., J. H. Stock, and M. W. Watson (2003). Macroeconomic forecasting in the euro area: Country specific versus area-wide information. *European Economic Review 47*, 1–18.

Marcellino, M., J. H. Stock, and M. W. Watson (2006a). A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of Econometrics 135*(1), 499–526.

Marcellino, M., J. H. Stock, and M. W. Watson (2006b). A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of Econometrics 135*(2), 499–526.

Masini, R. and M. C. Medeiros (2019). Counterfactual analysis with artificial controls: Inference, high dimensions and nonstationarity. Technical report, SSRN.

McCracken, M. W. and S. Ng (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*, 574–589.

Medeiros, M. C. and E. F. Mendes (2016). ℓ_1-regularization of high-dimensional time series models with non-gaussian and heteroskedastic errors. *Journal of Econometrics 191*, 255–271.

Meier, L., S. Van De Geer, and P. Bühlmann (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B 70*(1), 53–71.

Moon, H. R. and B. Perron (2012). Beyond panel unit root tests: Using multiple testing to determine the non stationarity properties of individual series in a panel. *Journal of Econometrics 169*(1), 29–33.

Müller, U. K. and G. Elliott (2003). Tests for unit roots and the initial condition. *Econometrica 71*(4), 1269–1286.

Nardi, Y. and A. Rinaldo (2011). Autoregressive process modeling via the lasso procedure. *Journal of Multivariate Analysis 102*, 529–549.

Ng, S. (2008). A simple test for nonstationarity in mixed panels. *Journal of Business and Economic Statistics 26*(1), 113–127.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics 92*(4), 1004–1016.

Onatski, A. and C. Wang (2018). Alternative asymptotics for cointegration tests in large VARs. *Econometrica 86*, 1465–1478.

Onatski, A. and C. Wang (2019). Extreme canonical correlations and high-dimensional cointegration analysis. *Journal of Econometrics 212*(1), 307 – 322.

Palm, F. C., S. Smeekes, and J.-P. Urbain (2008). Bootstrap unit root tests: comparison and extensions. *Journal of Time Series Analysis 29*(1), 371–401.

Palm, F. C., S. Smeekes, and J.-P. Urbain (2010). A sieve bootstrap test for cointegration in a conditional error correction model. *Econometric Theory 26*(3), 647–681.

Palm, F. C., S. Smeekes, and J.-P. Urbain (2011). Cross-sectional dependence robust block bootstrap panel unit root tests. *Journal of Econometrics 163*, 85–104.

Pantula, S. G. (1989). Testing for unit roots in time series data. *Econometric Theory 5*(2), 256–271.

Park, J. Y. and P. C. Phillips (1988). Statistical inference in regressions with integrated processes: Part 1. *Econometric Theory 4*(3), 468–497.

Park, J. Y. and P. C. B. Phillips (1989). Statistical inference in regressions with integrated processes: part 2. *Econometric Theory 5*, 95–131.

Park, T. and G. Casella (2008). The bayesian lasso. *Journal of the American Statistical Association 103*, 681–686.

Pedroni, P., T. J. Vogelsang, M. Wagner, and J. Westerlund (2015). Nonparametric rank tests for non-stationary panels. *Journal of Econometrics 185*(2), 378–391.

Pesaran, M. H., A. Pick, and A. Timmerman (2011). Variable selection, estimation and inference for multi-period forecasting problems. *Journal of Econometrics 164*, 173–187.

Phillips, P. C. and B. E. Hansen (1990). Statistical inference in instrumental variables regression with I(1) processes. *Review of Economic Studies 57*, 99–125.

Phillips, P. C. and S. Ouliaris (1990). Asymptotic properties of residual based tests for cointegration. *Econometrica 58*, 165–193.

Phillips, P. C. B. and V. Solo (1992). Asymptotics for linear processes. *Annals of Statistics 20*, 971–1001.

Rho, Y. and X. Shao (2019). Bootstrap-assisted unit root testing with piecewise locally stationary errors. *Econometric Theory 35*(1), 142–166.

Romano, J. P., A. M. Shaikh, and M. Wolf (2008a). Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test 17*(3), 417–442.

Romano, J. P., A. M. Shaikh, and M. Wolf (2008b). Formalized data snooping based on generalized error rates. *Econometric Theory 24*(2), 404–447.

Romano, J. P. and M. Wolf (2005). Stepwise multiple testing as formalized data snooping. *Econometrica 73*(4), 1237–1282.

Schiavoni, C., F. Palm, S. Smeekes, and J. van den Brakel (2019). A dynamic factor model approach to incorporate big data in state space models for official statistics. arXiv e-print 1901.11355, arXive.

Schwert, G. W. (1989). Tests for unit roots: a Monte Carlo investigation. *Journal of Business and Economic Statistics 7*(1), 147–159.

Shao, X. (2010). The dependent wild bootstrap. *Journal of the American Statistical Association 105*(489), 218–235.

Shen, H. and J. Z. Huang (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis 99*, 1015–1034.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics 22*(2), 231–245.

Smeekes, S. (2015). Bootstrap sequential tests to determine the order of integration of individual units in a time series panel. *Journal of Time Series Analysis 36*(3), 398–415.

Smeekes, S. and A. M. R. Taylor (2012). Bootstrap union tests for unit roots in the presence of nonstationary volatility. *Econometric Theory 28*(2), 422–456.

Smeekes, S. and J. Urbain (2014a). A multivariate invariance principle for modified wild bootstrap methods with an application to unit root testing. GSBE Research Memorandum RM/14/008, Maastricht University.

Smeekes, S. and J.-P. Urbain (2014b). A multivariate invariance principle for modified wild bootstrap methods with an application to unit root testing. GSBE Research Memorandum RM/14/008, Maastricht University.

Smeekes, S. and J.-P. Urbain (2014c). On the applicability of the sieve bootstrap in time series panels. *Oxford Bulletin of Economics and Statistics 76*(1), 139–151.

Smeekes, S. and E. Wijler (2018a). An automated approach towards sparse single-equation cointegration modelling. arXiv e-print 1809.08889, arXive.

Smeekes, S. and E. Wijler (2018b). Macroeconomic forecasting using penalized regression methods. *International Journal of Forecasting 34*(3), 408–430.

Smeekes, S. and E. Wijler (2020). Unit roots and cointegration. In P. Fuleky (Ed.), *Macroeconomic Forecasting in the Era of Big Data*, Volume 52 of *Advanced Studies in Theoretical and Applied Econometrics*, Chapter 17, pp. 541–584. Springer.

Song, S. and P. J. Bickel (2011). Large vector auto regressions. Technical Report arXiv:1106.3915, arXiv.

Stock, J. H. and M. W. Watson (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*, 1167–1179.

Stock, J. H. and M. W. Watson (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics 20*, 147–162.

Stock, J. H. and M. W. Watson (2003). Has the business cycle changed and why? In M. Gertler and K. Rogoff (Eds.), *NBER Macroeconomics Annual 2002, Volume 17*, Chapter 4, pp. 159–230. MIT Press.

Stock, J. H. and M. W. Watson (2006). Forecasting with many predictors. *Handbook of Economic Forecasting 1*, 515–554.

Stock, J. H. and M. W. Watson (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics 30*, 481–493.

Strassen, V. (1967). Almost sure behavior of sums of independent random variables and martingales. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Contributions to Probability Theory, Part 1*, pp. 315–343. University of California Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*, 267–288.

Trapani, L. (2013). On bootstrapping panel factor series. *Journal of Econometrics 172*(1), 127–141.

Trapletti, A. and K. Hornik (2018). *tseries: Time Series Analysis and Computational Finance*. R package. R package version 0.10-46.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics 12*, 389–434.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning 8*, 1–230.

Van de Geer, S. (2007). The deterministic lasso. Seminar proceedings, Eidgenössische Technische Hochschule (ETH) Zürich.

Van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics 42*, 1166–1202.

Van De Geer, S. A. and P. Bühlmann (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics 3*, 1360–1392.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, Volume 47. Cambridge University Press.

Wagener, J. and H. Dette (2013). The adaptive lasso in high-dimensional sparse heteroscedastic models. *Mathematical Methods of Statistics 22*, 137–154.

Wang, H., G. Li, and C. Tsai (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of The Royal Statistical Society B 69*, 63–78.

Westerlund, J. (2007). Testing for error correction in panel data. *Oxford Bulletin of Economics and statistics 69*(6), 709–748.

Wilms, I. and C. Croux (2016). Forecasting using sparse cointegration. *International Journal of Forecasting 32*, 1256–1267.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences 102*, 14150–14154.

Yamada, H. (2017). The frisch–waugh–lovell theorem for the lasso and the ridge regression. *Communications in Statistics-Theory and Methods 46*(21), 10897–10902.

Yoon, Y. J., C. Park, and T. Lee (2013). Penalized regression models with autoregressive error terms. *Journal of Statistical Computation and Simulation 83*, 1756–1772.

Zhang, R., P. Robinson, and Q. Yao (2019a). Identifying cointegration by eigenanalysis. *Journal of the American Statistical Association 114*(526), 916–927.

Zhang, R., P. Robinson, and Q. Yao (2019b). Identifying cointegration by eigenanalysis. *Journal of the American Statistical Association 114*, 916–927.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine Learning Research 7*, 2541–2563.

Ziel, F. (2016). Iteratively reweighted adaptive lasso for conditional heteroscedastic time series with applications to AR-ARCH type processes. *Computational Statistics and Data Analysis 100*, 773–793.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics 15*, 265–286.

Valorization

The central theme underlying this thesis is the analysis of high-dimensional time series datasets. The current era is characterized by wide availability of larger and less structured datasets and the process of information extraction from this kind of data demands a drastically different approach that better accommodates these new features. While much progress is being made in the field of high-dimensional statistics in recent years, the analysis of high-dimensional time series in particular merits additional treatment. The appeal of high-dimensional time series analysis stems from the idea of drawing strength from both the time dimension and the (potentially large) cross-sectional dimension to improve model estimates and corresponding forecasts. However, time series analysis in high dimensions comes with its own unique set of challenges. First, the accelerated growth in time series datasets is experienced mostly along the cross-sectional dimension, as changes in information management allow us to extract data from more individuals or agents, but the passing of time puts a strict limit on the growth in the time dimension. Second, even on traditional, smaller datasets the peculiarities of time series such as serial dependencies, non-stationarity and structural breaks call for specialized treatment. The addition of high-dimensionality exacerbates the complexity of the analysis of time series, and the research presented in this thesis aims to contribute to this problem in several ways.

A strong emphasis in this thesis is placed on rigorous and, especially, honest comparison between traditional and state-of-the-art statistical models that have a strong founding in econometric theory. As often the case in transitional periods, it is easy to become convinced by ill-founded claims or idiosyncratic success stories of new methods in exotic applications. Accordingly, the second chapter consolidates seminal and recent literature on prospective statistical methods that are well-suited for forecasting based on high-dimensional time series datasets, and contains elaborate

comparisons of their forecast performance in both controlled and real-life settings. The results provide detailed insights into the relationship between the considered estimators' forecast performance and characteristics of the data (generating process). These insights can serve as a guideline for practitioners facing a forecasting exercise or provide useful benchmarks for the development of new estimators. Furthermore, the value in this chapter is strengthened by our focus on general and realistic data characteristics that are not necessarily specific to a particular field, thus allowing for broad applicability. On a personal level, I particularly hope the research presented in this chapter can be of use to the field of climate science, where temperature forecasts are often based on large datasets of atmospheric measurements containing time series that are characterized by strong (seasonal) dependence over time and cross-sectional dependence due to the proximity between measuring stations. Liberally conjecturing on potential applications, I consider (1) the use of penalized regression to filter out irrelevant atmospheric particles types from the data, (2) using principal component based algorithms to impute missing or faulty measurements and (3) modelling large cointegrated systems of, for example, land and sea temperatures combined with greenhouse gasses as interesting avenues of research that the results in this thesis may be able to contribute towards.

In the third and fourth chapter we develop the Single-equation Penalized Error Correction Selector (SPECS), a novel estimator that combines the traditional approach of cointegration modelling in conditional systems with the dimensionality reduction properties of penalized regression. Ever since its development, cointegration modelling has been an essential tool in the study of economic relationships, with classical examples including purchasing power parity (Juselius and MacDonald, 2004), money demand (Johansen, 1992b) and rational expectation models (Johansen and Swensen, 2004), as well as the study of financial theory such as the present value model of stock prices (Campbell and Shiller, 1987), market efficiency (Dwyer Jr and Wallace, 1992) and numerous market linkages such as that between local gasoline prices and global oil prices (Hendry and Juselius, 2001). More modern applications examine these kind of phenomena on a global scale based on cointegrated panel data (e.g. Westerlund, 2007), where the large cross-sectional dimension calls for specialized high-dimensional methods. If in this cases, the modelling exercise is focussed around only a few variables of interest, SPECS can serve as an automated tool to fit sparse linear single-equation models that incorporate both the long-run and short-run dynamics in the data. As demonstrated in the empirical application of Chaper 3, in which we nowcast Dutch unemployment based on Google Trends data, SPECS is particularly well-suited for the purpose of nowcasting economic variables on such

large macro-economic datasets. Buono et al. (2017) provide an interesting survey of recent studies incorporating various novel sources of big data, such as Google Trends, credit card, social media and stock exchange data, to nowcast macro-economic variables. With the rise of this many sources of big data, the nowcasting potential of SPECS has clearly not been exhausted in the single empirical application considered in Chapter 3.

SPECS may also be used for so-called "Artifical Counterfactual Analysis" (ArCo) in the spirit of Masini and Medeiros (2019). Counterfactual analysis is the examination of treatment effects in the absence of obvious control groups. For example, the highest income tax in the Netherlands was changed in 2001 from 60% to 51%. A natural question to ask is how this has impacted the GDP of the Netherlands. To disentangle the effect of the tax law change and other variables affecting the DGP, one may consider the use of neighbouring economies that were not subjected to this policy change as artificial control groups. Creating artificial controls based on multiple countries and multiple economic indicators quickly gives rise to high-dimensional models, for which SPECS can be considered as a useful estimator. While some theoretical details ought to be worked out, the estimation consistency of SPECS derived in the high-dimensional framework of Chapter 4 is a valuable pre-requisite for ArCo based on SPECS to be considered valid.

For many statical models that form the basis for the determination of economic policy, the ability to perform honest, i.e. uniformly valid, post-selection inference on large (co)integrated datasets is essential. I acknowledge that the thesis does not contribute to this important topic directly. However, from the post-selection inferential tools developed in the stationary world, such as the desparsified lasso (Van de Geer et al., 2014) or post-double selection method (Belloni et al., 2014), it is clear that the theoretical results derived in Chapters 3 and 4 may serve as starting points for the development of novel inferential techniques.

It is worth mentioning that, from the start of the development of SPECS, key considerations have been the intuitiveness of the model and ease of implementation. I believe that the value of an estimator is ultimately derived from its practical usability and the adoption rate among practitioners, no matter how mathematically interesting the underlying theory may be. Not only do I believe that the resulting single-equation model is understandable for a wide audience including non-experts, it is implementable with readily available, off-the-shelf tools including self-written code that I have made publicly available online. Moreover, the relatively low requirements in terms of data pre-processing further reduces the burden on the applied researcher.

In light of the results in Chapter 5 that demonstrate the complexity of controlling (family-wise) error rates of unit root tests in high-dimensions, I consider this automation of the model building process particularly valuable.

Finally, I would like to take the liberty of including an element that is not traditionally part of the valorisation of a thesis: teaching. The accumulation of knowledge throughout a PhD would be worthless to society without its subsequent dissemination. The publication of scientific results tends to reach a rather select audience, whereas knowledge transfer through direct interaction with students often has much farther reaching consequences. Having been lost on my academic path for a while myself, I understand the value of guiding young students in their search for knowledge and self-development. I have had the fortune to teach students from all over the world, with equally varying backgrounds, and made it my goal to connect with them and to convince them of the value of quantitative analysis. Of course, teaching an already excited econometrics student about the power and generality of maximum likelihood estimation has been a great pleasure, but witnessing social science students discovering the value of statistical inference within their fields of interest and, often to their own surprise, becoming excited about statistical theory, felt equally rewarding. I hope I have achieved my goals of inspiring the new generation to pursue their academic interests and I look forward to what the future may bring.

Nederlandse Samenvatting

We bevinden ons momenteel in een nieuw tijdperk van data-analyse, dat gekarakteriseerd wordt door de beschikbaarheid van grote, ongestructureerde datasets. U kunt hierbij denken aan data die wordt verzameld door grote tech-bedrijven zoals Google en Facebook, maar ook gegevens die verzameld worden via de klantenkaart van de lokale supermarket en de betaalpas waarmee afgerekend wordt. Omdat traditionele statistische modellen vaak het beste werken wanneer er rekening gehouden dient te worden met de effecten van *slechts enkele* variabelen, zijn er de laatste jaren veel nieuwe statistische methoden ontwikkelt die beter toepasbaar zijn op grote datasets. Deze nieuwe methoden worden ook wel hoog-dimensionale statistieken genoemd. Echter, binnen economische en financiële sectoren, werkt men met name met tijdreeksen, zoals bijvoorbeeld de Nederlandse werkloosheidcijfers of het bruto binnenlands product. Tijdreeksen vertonen vaak unieke eigenschappen, zoals trendmatig gedrag waarbij toekomstige waardes sterk afhangen van het verleden, waarvan we weten dat ze de uitkomsten van traditionele statistieken sterk beïnvloeden. Het is daarom niet verstandig om hoog-dimensionale statistieken toe te passen op grote verzamelingen van tijdreeksen zonder theoretische verificatie of praktische aanpassingen. Dit onderwerp staat centraal in mijn proefschrift.

In dit proefschrift, richten we ons enkel op statistische methoden welke onder te verdelen zijn in drie algemene categorieën: (1) factor modellen, (2) geregulariseerde regressie en (3) hybride modellen. Het idee achter factormodellen is dat alle waargenomen variabelen worden aangedreven door enkele latente (niet geobserveerde) variabelen. Zo kunnen we bijvoorbeeld werkloosheid observeren binnen verschillende industrieën, of rentetarieven voor verschillende looptijden, maar worden al deze variabelen mogelijk (deels) verklaard door de onderliggende bedrijfsconjunctuur. Factor modellen proberen deze latente variabelen, de factoren, te schatten en daarmee de

data samen te vatten met een minimum verlies aan informatie. Op deze manier hoeft er geen complex model met honderden geobserveerde variabelen geschat te worden. Een alternatieve methode is om de data niet samen te vatten, maar om ervan uit te gaan dat veel variabelen simpelweg irrelevant zijn voor het verklaren van de afhankelijke variabele waar men in geïnteresseerd is. Zo is het aannemelijk dat de grondstofprijzen voor thee van invloed zijn op de verkoop van koffie, maar dat de grondstofprijzen voor ketchup hier weining in verklaren. Voor dit soort applicaties is geregulariseerde regressie uitermate geschikt. Deze vorm van regressie schat een lineair model en zorgt er automatisch voor dat de geschatte bijdrages van irrelevante variabelen omlaag geschaald worden. Sommige vormen van geregulariseerde regressie, zoals de Least Absolute Shrinkage and Selection Operator (LASSO) welke een belangrijke rol in dit proefschrift heeft, hebben de wenselijke eigenschap dat ze irrelevante variabelen geheel automatisch uit het geschatte model kunnen verwijderen. Als laatste optie komen in dit proefschrift hybride methoden aan bod, welke irrelevante variabelen verwijderen en de relevante variabelen middels het schatten van factoren samenvatten.

In Hoofdstuk 2 vergelijken we de voorspellingsprestaties van statistische methoden welke onder te verdelen zijn middels de bovenstaande categorisatie. Door het uitvoeren van gecontroleerde simulaties waarin we bepaalde data eigenschappen doelbewust vastleggen, vinden we dat factor modellen en geregulariseerde regressie goed presteren in het kader waar ze voor ontwikkeld zijn, maar ontdekken we ook dat geregulariseerde regressie beter kan voorspellen indien er factoren in de data aanwezig zijn met "veel ruis".[1] In een empirische toepassing vinden we dan ook dat voor sommige Amerikaanse economische indicatoren geregulariseerde regressie nauwkeuriger voorspelt dan factor modellen, ondanks dat de aanwezigheid van factoren in een macro-economische toepassing zeer aannemelijk is.

Gemotiveerd door de gunstige prestaties van geregulariseerde regressie, ontwikkelen we in Hoofdstuk 3 de Single-equation Penalized Error-Correction Selector (SPECS). SPECS is een gespecializeerde methode waarmee geregulariseerde lineaire modellen geschat kunnen worden die rekening houden met het trendmatige gedrag van de beschouwde variabelen. Zo komt het in economische toepassingen geregeld voor dat individuele variabelen een stochastische (willekeurige) trend bevatten, maar dat deze trend verdwijnt na het nemen van een bepaalde lineaire combinatie. Dit welbekende fenomeen heet cointegratie en heeft grote invloed op het gedrag van statistieken. Wij

[1] Dit is een simplificatie ter bevording van de leesbaarheid. De preciezere omschrijving is dat cross-sectionele correlatie in het idiosyncratische component de nauwkeurige schatting van factoren belemmert.

leiden theoretische (asymptotische) resultaten af die laten zien dat onze methode zich wenselijk gedraagt wanneer de steekproefgrootte groeit. Ter demonstratie van de toepasbaarheid van SPECS, gebruiken we onze nieuwe methode om de werkloosheid in Nederland te voorspellen aan de hand van de populariteit van 100 verschillende Google zoektermen, waaronder bijvoorbeeld "werkloosheidsuitkering" en "solliciteren". In lijn der verwachtingen, overtreft SPECS de voorspellingsprestaties van hoog-dimensionale statistieken welke cointegratie negeren.

In Hoofdstuk 4 leiden we vergelijkbare theoretische resultaten af onder minder restrictieve aannames. Zo laten we toe dat het aantal variabelen in het model mag toenemen wanneer de steekproefgrootte toeneemt. Dit is van belang om een duidelijk inzicht te geven in het gedrag van SPECS bij toepassingen op datasets met een groot aantal variabelen.

Ten slotte, in Hoofdstuk 5 vergelijken we (1) statistische testen om het trendmatig gedrag van tijdreeksen te classiferen en (2) een selectie aan hoog-dimensionale voorspellingsmethoden welke cointegratie al dan niet in acht nemen. Middels simulaties vinden we dat het uitermate belangrijk is om de trend in de afhankelijke variabele juist te classificeren, gezien de nauwkeurigheid waarmee deze variabele voorspeld kan worden sterk van deze classificatie afhangt. In een macro-economische toepassing op een Amerikaanse dataset vinden we dat geen enkel model consistent het nauwkeurigst voorspelt en is er ook geen definitief antwoord op de vraag of cointegratie belangrijk is voor het maken van voorspellingen. Gezien er gevallen zijn waarin SPECS beter presteert dan de andere methodes in de vergelijking, bevestigen we dat onze methode zowel theoretische als toegepaste waarde heeft. Echter, zal de keuze voor de optimale methode altijd van de specifieke toepassing afhankelijk zijn.

Curriculum Vitae

Etienne Wijler was born on March 20, 1991 in Maastricht, The Netherlands. Between 2005 and 2009 he attended high school at Bonnefanten College in Maastricht. His education continued with a bachelors degree in international business at the Zuyd University of Applied Sciences. After obtaining his bachelor degree in 2013 (cum laude), he started his masters at Maastricht University in international business with a specialization in finance. He wrote his master thesis on 'The impact of rating differentials on loan spreads' and graduated cum laude in 2015. While still being enrolled in the finance track, Etienne started a second masters in econometrics and operations research at Maastricht University. He particularly enjoyed the econometrics component of the program and his master thesis entitled 'Comparing the prediction accuracy of high-dimensional statistics' marked the successful completion of his second master degree. After this graduation, Etienne started his PhD. under the supervision of the late Prof. dr. Jean-Pierre Urbain and dr. Stephan Smeekes. The supervisory tasks from Prof. dr. Jean-Pierre Urbain were later taken over by Prof. dr. Alain Hecq. The findings of this research are presented in this dissertation.

In January 2020 Etienne started a post-doctoral fellowship at Maastricht University, continuing his research in high-dimensional time series analysis.