Inference in high-dimensional time series models

Inference in High-Dimensional Time Series Models

Luca Margaritella

This research was financially supported by the Netherlands Organization for Scientific Research (NWO) under grant number 452-17-010.

This book was typeset by the author using LaTeX

Published by Universitaire Pers Maastricht
ISBN: 978-94-6423-510-4
Printed in The Netherlands by ProefschriftMaken

Inference in High-Dimensional Time Series Models

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 Monday 8th November 2021, at 10:00 hours

by

Luca Margaritella

Supervisors:

Dr. Stephan Smeekes

Prof. dr. Alain Hecq

Assessment Committee:

Prof. dr. Jaap Bos (chair)

Dr. Anders Bredahl Kock

Prof. dr. Claudio Morana

Dr. Ines Wilms

To Margherita & Nicolò

Acknowledgments

Looking back to 4 years ago today, one thing seems clear to me: that I have been outrageously lucky. I will try to explain why in the following lines but let me start by saying that I am extremely grateful. As time series are the "leitmotif" of what comes in the next chapters of this dissertation, let me pick an annual frequency t. While I am writing this we are in $t = 2021$, therefore let me start with $t - 5 = 2016$.

By that time I was fresh out of my bachelor in Statistics. By far what I enjoyed the most during those years in Milan were the courses of Econometrics and Time Series Analysis. I recall I approached Prof. Claudio Morana as I thought I had some mildly cool results out of my BSc thesis. In restrospect, I would say it was quite garbage but Claudio was nice enough to look at it and give me feedbacks even if I was not his student. Furthermore, Claudio involved me in the organizational machine for the annual conference of the International Association for Applied Econometrics (IAAE), which that year was held in Milan. It was at the conference dinner that Claudio insisted for me to meet someone he knew from Maastricht Universty that was sitting at another table. Little that I knew, that meeting shaped the following five years of my life. Here I must immediately add that I am of course incredibly grateful for all the help and support Claudio spontaneously gave me. The person he introduced me would have become my supervisor for the MSc and co-supervisor for my PhD. Alain (more about him later) was in fact sitting at that table. Sipping some red wine (cannot blame him), Alain listened to my confused ramblings about how interested I was in econometrics and time series models. Convinced by Claudio and the talk with Alain which continued via emails, I decided in $t = 2016, 2017$ to enrol in a master in Econometrics and Operation Research at Maastricht University. My year as a master student was not easy (let alone digesting game theory): it took me some time to adjust to the style and pace used in teaching in Maastricht, quite different with respect to the italian system I was used to, but it was surely a very enriching experience. I learned a lot, took a lot of rain on my head going back and forth from the belgian place I was living to at the time, but I also met many friends, some of which later became PhD fellows with me.

Elisa, little french girl in appearence, party animal underneath: still memorable few parties at your place as well as at Adam's place which continued in the most questionable clubs of the city. Thank you both for having had always a bright spirit! Dewi, probably the first true Maastrichter I got to know. Misteriously able to party until crazy hours in the morning and get straight top grades in every course. I always enjoyed your company, your Dutch directness and the many chats we had along the way, thank you for the good times! Niels: apparently you were also there at the master, but where?! (more on Niels later).

The year $t = 2016$ has been a difficult one for the whole KE department. The passing of Jean-Pierre, which I unfortunately never got to know personally, has affected many. Nevertheless, the courses had to continue and the whole department has shown lot of strength and dedication. I managed to learn a lot that year and I especially treasure the many moments I've been able to talk with Alain by just dropping by his office. We've talked about many things: from the MSc thesis of course, to my plans for the future. I always felt extremely comfortable and always welcomed and I cannot thank enough Alain for all his help throughout the years. The relationship with Alain was truly a game-changer for me, used to useless academic hierarchies, confusing respect with lack of human empathy. He, and subsequently others I got to know better in the department, showed me that there exists a way of "doing academia" that can be really great.

The academic year $t = 2016, 2017$ has also been the first where a course on high-dimensional econometric methods was taught. To this day, whenever I have to teach something, I feel a big responsibility since I realize that basically all the choices I took in terms of research directions were driven by what I have been taught by some outstanding teachers in those very topics. Stephan is one of those. Hands-down, for me, that was the best course of the MSc and got me wildly interested in the topic. So interested, that when I saw that Stephan was looking for PhD students I did not really think twice before sending him an email. Among the very many (countless?) things for which I am grateful to Stephan, surely the first one is the trust he has put in me, taking me as his PhD student on this topic.

Indeed, in the academic year $t = 2017, 2018$ I started my PhD under the supervision of Stephan and Alain. The amount of things I learnt from the two of them, I don't even start with it. From pitching ideas, develop methodology and theory, come up with relevant empirical applications, write papers, present at seminars and conferences, do not insult referees and the list continues. I am very grateful to both of them and I consider myself very lucky that I got to know them and work with them. When other fellow PhDs were asking me: ”How is it going with your supervisors?”. I was always: ”Great! the two are a perfect combination. With both of them we shape the ideas and then with Stephan we take care of theoretical aspects and with Alain the applied aspects.” The formula has been great indeed. I would particularly like to say that I am also grateful for their great encouragement in presenting at conferences. This allowed me to, first of all, keep up with novelties in the field but also travel and, most of all, meet other academics. Several of the great coauthors I collaborate with today stemmed from these opportunities I had. I truly enjoyed (in pre-Covid times) all the conferences I have attended, even those where I had conceptual fights with big shot econometricians (ups!).

As I started my PhD, the “old guard” of PhD cohort has been super nice with me and I am thankful for having met all of them. Sean, Rasmus, Alex, Yicong, Tim, Veerle, Roland, Hanno, Etienne. Thanks for having been so welcoming and nice with me. Special thanks to the 'metrics crew: Hanno and Etienne, I am really greatful for the many times you've been available to chat about my research topics or suggesting R tricks. Sean, Rasmus, Yicong and Alex: I always looked up at you as great and talented researchers. Spending time with you all guys, also outside the uni, either cooking or roaming the city was really fun!

My PhD years $t = 2018, 2019, 2020, 2021$ would have however not been the same if not for the amazing other PhDs that started the same year as me. In particular, let me start from my office mate Benoit. As I haven't had the opportunity to address him in length at his (on-line) defense, I am going to do this here. These years have been a true blessing for anyone coming in contact with Benoit at KE and outside. For me, especially, it has been extra-special as I got to share the A4.25 office -our office- with Benoit for three (too quick) years. Looking back in time, I recall our initial chats in the morning before starting to try to understand what to do with research. Coffee was bad (still is), no staff-lounge was down there back then, but this has never stopped us from getting a warm cup together. Ben even brought his filter-coffee machine at some point, along with some biological coffee, that was even worse but I drank it nonetheless, it did not matter. During our initial conversations I remember that to overcome some English blanks Ben had initially, we were either using French words or Italian words, at the end of the day that was easier to pass the concept. We have passed long hours in the office, it has been really the extension of our respective homes and in that sense I always felt I had a housemate and a friend close to me. Ben's passion and talent for math are "obvious" to anyone knowing him well enough and he has done an outstanding job during these years as a PhD candidate. Even though econometrics is not his expertise, he helped me so many times, chatting over the whiteboard about the theory of my papers, I am very thankful for that. Soon enough I started to invade the tiny available spaces in the office with plants and I'm still glad Ben was happy with it, because I must admit I have been annoying with all that gardening sometimes in the morning. Our friendship rapidly grew in time, Ben has always been an incredible aggregator: I think the set of groups on WhatsApp he created for hiking, cooking, movies, beers, dinners, Mario-kart, you name it, is surely countably infinite. Ben's house, as a matter of fact, became sort of the head-quarter for many of us PhDs+ at KE. In that little, barely 25 square meters, we fit in up to 12 people with beers, food, during movies projected on the wall, to just chat and complain about teaching and students over few sometimes funny-tasting dry fruits, nuts, lemon pies, rhubarb pie, home-made limoncello, a never ending selection of Belgian beers, so strong that after one you felt dizzy and question your ability to drive the bike back home. Ben: you have always been extremely happy of all this and I have been likewise, with a feeling of admiration on how self-giving you have always been. These years have been a life experience difficult to forget and a really big part of this is because of you Ben! Thank you!

Niels: while I was unaware of you in the master, I'm surely happy we got to know each others well during the PhD! I cherish the many hikes you expertly organized where we could conveniently end up in Noorbeek at the cafè with the best onion soup and vlaai ever witnessed! The many evenings at Benoit's place for movies or movie marathons. The trips together: Berlin with its distinct piss-smell and the random guy in the bed above you in the hostel that did not let you sleep. Biarritz, with the amazing sangria of Ben's mom. Prague, where we both conveniently got fever which however did not stop us from visiting all the beer gardens possible. More recently in Italy, where we never managed to eat in that restaurant where we originally planned and where was easier to solve the Goldbach's conjecture rather than finding a free parking spot.

Adi: I always admired your chilled spirit and your curiosity! When I think about someone that is never worried about anything, I think of you. We've sampled so many restaurants and pubs in Maastricht and found our "usual ones" where often times, after work, we were heading to. So many dinners at Ginger and pizza's at piano B! We shared this passion for whiskey and La Chouffe and we agreed on the obvious fact that at Tribunal La Chouffe tastes better than everywhere else. Thanks to you and Shash for introducing me to the proper Indian food and for spicing up (!) my flat European tastes. Especially, thanks for introducing me to bakarwadi, I cannot stop thinking about them, too good!

Memories are so many for few pages to put in. But memories are so great to remain in all of us.

Many others in the department I am grateful to have met:

Cate & Li, me and Benoit's office front-neighbors. Thank you both for your bright spirit and the many occasions we managed to share. Li, you're the least Chinese person among the Chinese people I know, it was a lot of fun to be around you! I still recall our first trip to a conference in Amsterdam when you brought me to a typical Chinese restaurant. I will hardly forget the spicy food you ordered for both of us and how sweaty I was and how steamy my glasses were already half way into the dish. I also recall your famous coca-cola chicken that you cooked for me once, it was really good! I'm still puzzled about the use of coca-cola in cooking though.

Niloufar & Farzaneh: the Iranians conquering the Netherlands! It was really fun to get to know you and spend time together traveling in Europe!

Aida: you almost convinced me that Eindhoven is a nice city. Okay, maybe not. But I'm very glad we got to know each other in time! It's been super fun whenever you were dropping by Maastricht to go out for dinners with all the others!

Julian, Francesco, Marie & Robert, I've really enjoyed talking with all of you about research and life.

Johannes & Thomas, memorable has been our trip to Cyprus for the IAAE (aka AIAIAI) conference. 40 degrees in the shadow, airco on 24/7 to avoid a certain death and our pilot Johannes driving on the right side through the island.

The Oktoberfest at Johannes' place in Munich, along with Kim and Mariana, has also been lots of fun (and too much beer!).

All the senior staff at UM as well as our great secretaries Karin, Yolanda and Vera: thank you all, I always felt absolutely at home in the department. Special thanks to Janòs, Ines, Denis, Nalan, Rui and many others.

Outside the university: Laura, thank you for all the good times and for the help with translating and navigating Dutch bureaucracies! Kamil, my Polish friend! We had great times in Maastricht as students and I'm so glad that these days we are able to meet in different parts of Europe to spend some time together!

My Italian crew of friends in Milan: Luca "Lux", Andrea "Andre", Valentina "M'bare", Silvia "Silvy", Fulvio "Fuz", Alessandro "Gigi", Marco "Crive", Filippo "Vèz", Leonardo "Leuccio". Thank you all guys, knowing to have so many friends back in Italy warms my heart and whenever I come back to Milan I know you guys are there for me and this definitely makes me feel extremely grateful.

Last, but certainly not least, I would like to express my utmost gratitude to my mom, Margherita, my brother Nicolò and my grandpa Alberto for their constant love, care and support throughout these years. My mom has always been and is the greatest source of love and support for me. Nic, blood brothers as well as stats brothers, thank you for your support in every choice I had to make in life. Nonno, one cannot forget the many years you dedicated to play with and take care of me. The fun and laughter we still have to these days are precious moments.

And here we are, at $t = 2021$, the PhD years have now passed and I carry with me so many beautiful moments and plenty of gratitude for you all. Now a new great adventure has started for me in Sweden. I am ready for the $t+1, t+2, t+3, \ldots$

Luca Margaritella
Lund
October 18, 2021

Contents

1 Introduction

This chapter is meant to introduce the reader with a set of topics that are central to the research developed in the following chapters. Chapter 2, 3, 4 of this dissertation focus on inference in high-dimensional time series models and especially on testing for Granger causality. Chapter 5 also deals with high-dimensional time series models, combining the two main schools of thoughts with what pertains dimensionality reduction, namely sparse and dense modeling.

Section 1.1 presents high-dimensional models and defines one of the central topics of this dissertation, namely post-selection inference. Section 1.1.1 first introduces sparsity-inducing techniques, specifically the penalized regression framework and how variable selection is attained using specific ℓ_q-norm penalties. Selection consistency and oracle properties are also presented. Furthermore, it also introduces the dense framework of factor models, underlines the differences with the sparse counterpart and sketches how factors are estimated via principal components. Section 1.1.2 addresses the main challenges to face when doing post-model selection inference, from an asymptotic perspective. How to obtain "honest inference" is discussed and one of the main approaches is outlined, namely the post-double selection. A formal treatment of the Granger causality concept is presented in Section 1.2 where some history of the concept is provided along with its formalization and a discussion of its problems. Section 1.3 outlines the contributions of this dissertation chapter-by-chapter.

1.1 High-Dimensional Models and Post-Selection Inference

1.1.1 High-Dimensional Models: Sparse and Dense

Historically statistics has dealt with low-dimensional settings where the number of observations in a data set, the sample size, is much greater than the number of variables, the features[1]. However, the technical advancements of the last twenty years have brought forward unprecedented possibilities in terms of data availability. Therefore, dealing with increasingly large data sets has become common practice both in academia and industry. Those data sets containing more features than observations are referred to as "high-dimensional". The aim of the present thesis is to develop statistical techniques for time series data, able to deal with such data sets. In fact, while data abundance offers great opportunities to describe and predict a variety of processes, from a statistical perspective it introduces several complications to deal with. Classical approaches in statistics such as linear regression, logistic regression etc., they are not suited for high-dimensional settings. For instance, when the amount of variables is as large as, or exceeds, the number of observations, then ordinary least squares (OLS) will return a set of coefficient estimates which perfectly fit the data, regardless of whether a true relationship exists between the features and the response. This is referred to as "overfitting": the perfectly fitted linear model does not prove any useful as the same model applied to an independent test set will yield very poor results. As a consequence, the variance of the (trained) model, namely its ability to generalize to other

[1] The terms features, covariates, regressors, predictors will be used interchangeably in the text.

test sets, will be large and lager than the bias counterpart[2]. Therefore, even though the OLS estimator is an unbiased estimator, the excessive variance will render its mean square error large, thus making the model perform very poorly in practice. Intuition would suggest that as the number of features used to fit a model increases, the quality of the fitted model will increase as well. This is not quite true as this depends on whether the additional features are truly relevant or just noise with respect to the response. Even in the unlikely case that only relevant features gets added to the model, the bias reduction derived from these additional features could be outweighted by the large variance incurred in estimating their coefficients. The issue lies in the fact that, while the parameter space grows at fast speed, its elements to estimate soon start to be too many for the sample information available to reliably estimate them, what is referred to as the "curse of dimensionality". If no additional structure is imposed on the model, specifically to the unknown regression vector, then there is no hope of obtaining consistent estimators when the ratio between the number of features and the sample size, stays bounded away from zero. To tackle this challenge, statisticians and econometricians have developed strategies which can be divided into two broad categories imposing a substantially different type of structure to the regression vector: sparse and dense modeling. Before introducing some of the details of the two philosophies, a mathematical formulation of the problem is presented.

Define $\boldsymbol{\beta} \in \mathbb{R}^d$ the unknown regression vector and suppose to observe a vector $\boldsymbol{y} \in \mathbb{R}^T$ and a matrix $\mathbf{X} \in \mathbb{R}^{T \times d}$. For instance, think of $\boldsymbol{y}$ as a time series of sample size T and $\boldsymbol{X}$ as a matrix containing d other time series of same length T. A linear model to link these variables is

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

[2] Recall the mean squared error (MSE) of an estimator $\hat{\beta}$ i.e., the measure of how well the estimator $\hat{\beta}$ is closed to the vector of parameters β, can be decomposed as $MSE(\hat{\beta}) = \mathbb{E}(\hat{\beta} - \beta)^2 = Var(\hat{\beta}) + [Bias(\hat{\beta})]^2$.

where $\boldsymbol{\varepsilon} \in \mathbb{R}^T$ is a vector of noise variables. For time series, the linear regression framework in (1.1) also encompasses seemingly unrelated regression models (SUR) such as vector autoregressive models (VARs) if for instance $\boldsymbol{X}$ contains the lags of $\boldsymbol{y}$ as well as lags of other covariates. Model (1.1) can also be written in a scalar form: for each index $t = 1, 2, \ldots, T$, one has $y_t = \langle x_t, \boldsymbol{\beta} \rangle + \varepsilon_t$, where $x_t^\top \in \mathbb{R}^d$ is the t-th row of $\mathbf{X}$, and y_t and ε_t are, respectively, the t-th entries of the vectors $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$. The quantity $\langle x_t, \boldsymbol{\beta} \rangle := \sum_{j=1}^{d} x_{tj}\beta_j$ denotes the usual Euclidean inner product between the vector $x_t \in \mathbb{R}^d$ of covariates and the regression vector $\boldsymbol{\beta} \in \mathbb{R}^d$. Thus, each response y_t is a noisy version of a linear combination of d covariates. In order to obtain a meaningful estimate of the regression vector $\boldsymbol{\beta}$, the model should have a low(er) dimensional structure. Assuming (strong) sparsity accomplishes this. In fact, by assuming the support set $S_{\boldsymbol{\beta}} := \{j \in 1, \ldots, d | \beta_j \neq 0\}$ of the regression vector β_j to have cardinality $s \ll d$ and the model being exactly supported on those s coefficients[3], there then exist techniques able to shrink the dimensions to only those relevant s coefficients. This variable selection is what penalized regression is set to accomplish. In the OLS framework, penalized regression techniques minimize the sum of squares residuals with an ℓ_q-norm penalty term added to the objective function. The ℓ_q-norm penalty term represents the constraint in the least squares minimization problem, avoiding the norm of the coefficient vector to become too large. This penalty term can be visualized by considering for a parameter $q \in [0, 1]$ and radius $r_q > 0$, the set

$$\mathbb{B}_q\left(r_q\right) = \left\{ \boldsymbol{\beta} \in \mathbb{R}^d \middle| \sum_{j=1}^{d} |\beta_j|^q \leq r_q \right\},$$

which is the set of ℓ_q-balls of radius r_q. According to the choice of q, these balls allow the estimates to be either shrunk towards zero but not exactly zero ($q > 1$) or shrunk towards zero and performing variable

[3] Note that assuming strong sparsity i.e., that the model is exactly supported on s coefficients may be overly restrictive. The notion of weak sparsity relaxes this: the vector $\boldsymbol{\beta}$ is weakly sparse if it can be closely approximated by a sparse vector.

selection by actually setting some coefficients equals to zero ($q \leq 1$). Figure 1.1 reports three instances for $d = 3$: (a) for $q = 1$, (b) for $q = 0.75$, (c) for $q = 0.5$.[4]

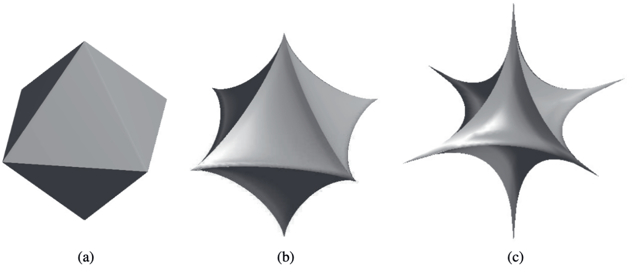


Figure 1.1: ℓ_q-Balls in $d = 3$: $(a) = \ell_1$, $(b) = \ell_{0.75}$, $(c) = \ell_{0.5}$

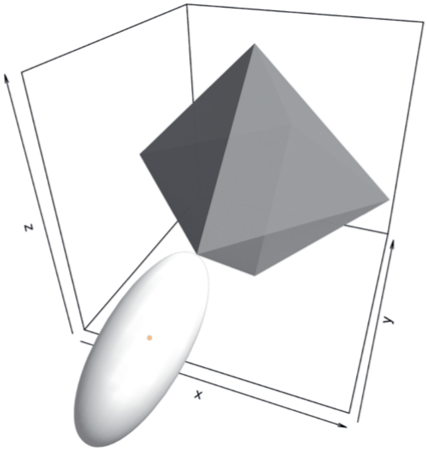
The penalized least square estimator is then obtained as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_q^q, \tag{1.2}$$

where λ controls the strength of the penalty: if large, then a strong shrinkage/variable selection is performed. If small, then in the limit the penalized least squares estimator approaches the OLS estimator. Note further, as clear from Figure 1.1, that only if $q \geq 1$ then the objective function is convex. The solution that minimizes the objective function (1.2) is located at the point where the (ellipsoid) contours of the sum of squared residuals cross the boundary of the constraint ℓ_q-ball as displayed in Figure 1.2, for the case of the ℓ_1-norm[5].

[4] The figure is taken from Wainwright (2019), Ch.7

[5] Figure 1.2 is taken from Stucky and Van De Geer (2017)



(a) ℓ_1-norm

Figure 1.2: Minimization solution with ℓ_1-norm

It follows that, as long as the shape of the ℓ_q-balls are sharp-cornered the solution is likely to lie at a corner point with one of the coefficients set equal to zero. A popular choice of q is indeed $q = 1$ which is referred to as the “lasso” which stands for “Least Absolute Shrinkage and Selection Operator” (Tibshirani, 1996). The lasso is able to combine shrinkage and variable selection with the convenience of a convex objective function[6]. In addition to the usual consistency argument for statistical estimators (i.e., $\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2 = o_p(1)$), all fitting procedure that combine simultaneous estimation and variable selection such as the lasso need that the set of relevant variables is correctly identified asymptotically with high probability[7]. This is referred to as selection

[6] Every local minimum will be a global minimum and hence only first order derivatives are needed. However, lasso is not differentiable and hence no analytical solution exists but one has to consider the subdifferential.

[7] A stronger consistency result than selection is sign consistency, (see Zhao and Yu,

consistency:

$$\mathbb{P}\left(\left\{j : \hat{\beta}_j \neq 0\right\} = \{j : \beta_j \neq 0\}\right) \to 1, \tag{1.3}$$

as $T \to \infty$. In addition, if under appropriate assumptions selection consistency holds for the lasso, then it follows that the variance of the estimated regression vector $\hat{\boldsymbol{\beta}}$ evaluated on the complement of the true support i.e., $\boldsymbol{\beta}_{S_\beta^c}$, is zero with high probability as $T \to \infty$. Therefore, the only relevant object for efficiency across different models will be the variance of the estimated regression vector on the true support i.e., $\boldsymbol{\beta}_{S_\beta}$. The so-called "oracle properties" have been introduced by Fan and Li (2001) to rank optimal fitting procedure that simultaneously attain variable selection and estimation. Intuitively, an oracle procedure will select the true set of relevant variables with probability one while estimating their coefficients as efficiently as if these relevant variables were known beforehand. Formally, $\hat{\boldsymbol{\beta}}$ is an oracle procedure if:

$$\begin{aligned} &(a) \quad \mathbb{P}\left(\hat{\boldsymbol{\beta}}_{S_\beta^c} = \mathbf{0}\right) \to 1, \\ &(b) \quad \sqrt{T}\left(\hat{\boldsymbol{\beta}}_{S_\beta} - \boldsymbol{\beta}_{S_\beta}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{S_\beta}\right), \end{aligned} \tag{1.4}$$

as $T \to \infty$, where $\boldsymbol{\Sigma}_{S_\beta}$ is the asymptotic covariance matrix of the OLS estimator on the variables constituting the true support. Lasso and its many refinements such as adaptive lasso (Zou, 2006), group lasso (Yuan and Lin, 2006), elastic net (Zou and Hastie, 2005) etc. have taken a substantial portion of the field's literature in the last 20 years. In all chapters of this thesis, the lasso is indeed the main character when it comes to dimensionality reduction.

Sparse dimensionality reduction techniques introduced thus far, impose a lower-dimensional structure to the regression vector by assuming some, or several, of its component to be irrelevant. As mentioned, this is not the only possible structure one can assume on $\boldsymbol{\beta}$. Factor models constitute an alternative way to attain dimensionality reduction.

2006)

They assume that the behavior of a certain variable can be decomposed into a component driven by few (r) unobservable (latent) factors ($\boldsymbol{F} = (F_1, \ldots, F_T)^\top$), which are common to the variables within a given data set but load differently ($\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_d)^\top$) on each of them, and a variable specific idiosyncratic component ($\boldsymbol{v} = (v_1, \ldots, v_T)^\top$). More formally, given model (1.1), assume the following decomposition holds:

$$\boldsymbol{X} = \boldsymbol{F}\boldsymbol{\Lambda}^\top + \boldsymbol{v}, \tag{1.5}$$

where $\boldsymbol{F}$ is the $T \times r$ matrix of common factors, $\boldsymbol{\Lambda}$ is the $d \times r$ matrix of factor loadings and $\boldsymbol{v}$ the $T \times d$ matrix of idiosyncratic components. Factors can be used for making predictions in place of $\boldsymbol{X}$, in fact by substituting $\boldsymbol{X} = \boldsymbol{F}\boldsymbol{\Lambda}^\top + \boldsymbol{v}$ in equation (1.1): $\boldsymbol{y} = (\boldsymbol{F}\boldsymbol{\Lambda}^\top + \boldsymbol{v})\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{F}\boldsymbol{\Lambda}^\top\boldsymbol{\beta} + \boldsymbol{v}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{F}\boldsymbol{\beta}_F + \tilde{\boldsymbol{\varepsilon}}$ for $\boldsymbol{\beta}_F := \boldsymbol{\Lambda}^\top\boldsymbol{\beta}$, $\tilde{\boldsymbol{\varepsilon}} := \boldsymbol{v}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. If the number of factors r is small and the factors approximate well the process $\boldsymbol{X}$, the dimensionality is greatly reduced with respect to the original column dimension of $\boldsymbol{X}$. Nevertheless, in general $\boldsymbol{F}$ is latent hence needs to be estimated from $\boldsymbol{X}$. The classical method employed to estimate the factors is principal components analysis (PCA). However, $\boldsymbol{F}$ is only identified up to rotation. In fact, taking an arbitrary $r \times r$ invertible matrix $\boldsymbol{H}$ such that $\boldsymbol{H}\boldsymbol{H}^{-1} = \boldsymbol{I}_r$ for $\boldsymbol{I}_r$ the identity matrix of order r, then it is immediate to observe how any model like $\boldsymbol{X} = \boldsymbol{F}\boldsymbol{\Lambda}^\top + \boldsymbol{v} = \boldsymbol{F}\boldsymbol{H}\boldsymbol{H}^{-1}\boldsymbol{\Lambda}^\top + \boldsymbol{v}$ could equivalently hold true. Hence, some identification restrictions are needed in order to have a unique $\boldsymbol{F}$: (i) $T^{-1}\hat{\boldsymbol{F}}\hat{\boldsymbol{F}}^\top = \boldsymbol{I}_r$, ($ii$) $\hat{\boldsymbol{\Lambda}}^\top\hat{\boldsymbol{\Lambda}}$ is diagonal. Under (i), (ii) the factor rotation is fixed and one can estimate factors and loadings using PCA. PCA minimizes the part of variance of $\boldsymbol{X}$ not explained by the factors. Formally, calling the columns of $\boldsymbol{X}$ as $\boldsymbol{X}_j$, similarly the columns of $\boldsymbol{v}$ as $\boldsymbol{v}_j$ and the transposed rows of $\boldsymbol{\Lambda}$ as $\boldsymbol{\Lambda}_j$ for $j = 1, \ldots, d$, then the PCA minimization problem can be written as:

$$(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\Lambda}}) = \underset{\boldsymbol{F},\boldsymbol{\Lambda}}{\arg\min}\, (Td)^{-1} \sum_{j=1}^{d} \|\boldsymbol{X}_j - \boldsymbol{F}\boldsymbol{\Lambda}_j\|_2^2. \tag{1.6}$$

The dimensionality reduction obtained via PCA estimation of the factors is such that the explained variability of the original set of variables is maximised given the number of factors, hence they are defined "dense models". Derivations of the PCA reveal how $\hat{\boldsymbol{F}}$ equals to the matrix of eigenvectors of $\boldsymbol{X}\boldsymbol{X}^\top$ corresponding to the largest r eigenvalues. Furthermore, factor models can be categorized as exact and approximate whether respectively the idiosyncratic components are assumed independent across the d variables i.e., $\mathbb{E}\left(\boldsymbol{v}_{t,j}, \boldsymbol{v}_{t',j'}\right) = 0$, $\forall t, t', j \neq j'$, or they are allowed to be weakly dependent i.e., $d^{-1}\sum_{j=1}^{d}\sum_{j'=1}^{d}\mathbb{E}(\boldsymbol{v}_{t,j}, \boldsymbol{v}_{t',j'}) < \infty$ as $d \to \infty$. Furthermore, dynamic is allowed to enter the factor model in terms of q lags of $\boldsymbol{F}$ in equation (1.5), thus distinguishing between static and dynamic factor models. However a dynamic factor model with q lags can also be written as a static factor model with $r(q+1)$ factors and hence estimated with PCA.

Both factor models and sparsity-inducing regression techniques are widely employed in practice and both have merits and shortcomings. The literature tends to polarize on either sparse or dense modeling. Chapter 5 reconciles the two factions retaining the best features of both by using a combination of a dynamic factor model where PCA is used in estimating the factors and a sparse VAR is used in estimating the idiosyncratic components.

1.1.2 High-Dimensional Models: the Problem of Inference

Let y_t, $x_{1,t}, \dots, x_{d,t}$ be a set of covariance-stationary[8] time series of interest for a sample size T and dimension d, potentially larger than T. Consider the following linear regression model

$$y_t = \sum_{j=1}^{d} \beta_j x_{j,t} + \epsilon_t = \boldsymbol{x}_t^\top \boldsymbol{\beta} + \epsilon_t, \quad t = 1, \dots, T, \tag{1.7}$$

[8] In Chapter 2 we work under this assumption, in Chapter 3 and 4 we relax it to consider unit root non-stationary time series as well.

where the intercept is omitted for simplicity, ϵ_t is a realization of a zero-mean stationary stochastic process and $\boldsymbol{\beta}$ is a d-dimensional vector of coefficients to estimate. Let

$$y_t = \beta_1 x_{1,t} + \boldsymbol{x}_{-1,t}^{\top} \boldsymbol{\beta}_{-1} + \epsilon_t, \tag{1.8}$$

where $x_{1,t}$ is a $T \times 1$ series of interest for testing a certain null hypothesis e.g., $H_0 : \ \beta_1 = 0$; and $\boldsymbol{x}_{-1,t}$ is the $T \times (d-1)$ matrix of potential confounders. As the high-dimensional setting is the focus, d is potentially larger than T, therefore one could use penalized regression techniques introduced in the earlier Section 1.1.1 to shrink the dimensionality of $\boldsymbol{x}_{-1,t}$. Ideally, one would think that by shrinking the dimension of $\boldsymbol{x}_{-1,t}$ and re-fitting the selected model with least squares would allow for a valid test of the relevant hypothesis on β_1. Inference on a model that has been selected from the data is called "post-model selection inference". However, the problem with it is that the model, being itself selected from the data, is random. The Oracle property in (1.9) ensures consistent model selection but this is not sufficient to have uniform convergence as the oracle property for post-selection estimators is a point-wise result, in other words, the estimator does not converge uniformly in the parameter space to a Gaussian distribution, but only point-wise. Point-wise limits can give very misleading results about approximations in finite samples. In order to illustrate this, consider the column dimension of $\boldsymbol{x}_{-1,t}$ to be just 1. Furthermore, consider for simplicity $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$, $\mathbb{E}(x_{1,t}\boldsymbol{x}_{-1,t}) \neq 0$ and that the oracle property in (1.9) holds for $\boldsymbol{\theta}_0 = (\beta_1, \boldsymbol{\beta}_{-1})^{\top}$; then $\forall \boldsymbol{\theta}_0$, $\|\boldsymbol{\theta}_0\|_2 \leq C$ for C a positive constant,

$$\lim_{T\to\infty} \mathbb{P}_{\boldsymbol{\theta}_0} \left(\hat{\boldsymbol{\beta}}_{-1,\hat{S}_\beta} = \boldsymbol{\beta}_{-1,S_\beta} \right) \to 1, \tag{1.9}$$

where $\hat{\boldsymbol{\beta}}_{-1}$ is the confounders vector $\boldsymbol{\beta}_{-1}$ estimated using a penalized regression method such as the lasso and $\hat{S}_\beta$ is the corresponding support. A famous negative result due to Leeb and Pötscher (2005) implies that

$\exists\delta > 0$:

$$\lim_{T\to\infty} \sup_{\|\boldsymbol{\theta}_0\|_2 \leq C} \mathbb{P}_{\boldsymbol{\theta}_0} \left(\underbrace{\left\| \sigma_\beta^{-1} \sqrt{T} (\hat{\beta}_1^{\text{OLS}} - \beta_1) - \mathcal{N}(0,1) \right\|_L > \delta}_{:=D} \right) \nrightarrow 0, \quad (1.10)$$

where $\hat{\beta}_1^{\text{OLS}}$ is the least squares estimator of β_1 after selection of $\hat{\boldsymbol{\beta}}_{-1}$, $\|g\|_L = \sup_{x\neq y} \frac{g(x)-g(y)}{|x-y|}$ for $g : [0,1] \to \mathbb{R}$, σ_β is the standard error of $\boldsymbol{\beta}_1$. The low dimensional setting well exemplifies this impossibility result. In fact, as the column dimension of $\boldsymbol{x}_{-1,t}$ is 1, there are actually only two model possibilities:

$$\begin{aligned} \alpha_I : \quad & \arg\min_{\beta_1} T^{-1} \|y_t - \beta_1 x_t\|_2^2; \\ \alpha_{II} : \quad & \arg\min_{\beta_1} T^{-1} \left\| y_t - \beta_1 x_t - \boldsymbol{x}_{-1,t}^\top \boldsymbol{\beta}_{-1} \right\|_2^2. \end{aligned} \quad (1.11)$$

The result in (1.10) is still true if one looks at a neighborhood of radius $\sqrt{T}$ of the true parameter vector i.e., if the supremum is taken over $\|\boldsymbol{\theta}_0\|_2 \leq C/\sqrt{T}$, for C a positive constant. Hence, fix e.g., $\beta_1 = 0$ and take a sequence for $\boldsymbol{\beta}_{-1,T} = \gamma/\sqrt{T}$ for γ a positive constant. Then the following lower bound is attained:

$$\begin{aligned} &\lim_{T\to\infty} \sup_{\|\boldsymbol{\theta}\|_2 \leq C/\sqrt{T}} \mathbb{P}_{\boldsymbol{\theta}_0}(D) \geq \lim_{T\to\infty} \sup_{\|\boldsymbol{\theta}\|_2 \leq C/\sqrt{T}} \mathbb{P}_{\left(0,\frac{\gamma}{\sqrt{T}}\right)}(D), \\ &\geq \lim_{T\to\infty} \mathbb{P} \left(\left\| \sigma_\beta^{-1} \sqrt{T} (\hat{\alpha}_I^{\text{OLS}}) - \mathcal{N}(0,1) \right\|_L > \delta \right) - \mathbb{P}_{\left(0,\frac{\gamma}{\sqrt{T}}\right)} \left(\hat{\boldsymbol{\beta}}_{-1,T}^{\text{OLS}} \neq 0 \right). \end{aligned} \quad (1.12)$$

However, the right hand side in the limit converges to $\mathbb{P}_{(0,0)} \left(\hat{\boldsymbol{\beta}}_{-1,T}^{\text{OLS}} \neq 0 \right)$ as $\gamma/\sqrt{T} \to 0$ as $T \to \infty$, which by the Oracle property in (1.9) is equal to zero. Hence, what remains is

$$\lim_{T\to\infty} \mathbb{P} \left(\left\| \sigma_\beta^{-1} \sqrt{T} (\hat{\alpha}_I^{\text{OLS}}) - \mathcal{N}(0,1) \right\|_L > \delta \right).$$

Now, the least squares estimator of model α_I ($\hat{\alpha}_I^{\text{OLS}}$) can easily be computed from (1.11):

$$\hat{\alpha}_I^{\text{OLS}} = \left(\frac{x_t^\top x_t}{T}\right)^{-1} \frac{x_t^\top y_t}{T}, \tag{1.13}$$

hence

$$\begin{aligned} &\sqrt{T}\left\{\left(\frac{x_t^\top x_t}{T}\right)^{-1} \frac{x_t^\top y_t}{T} - \beta_1\right\} = \\ &\sqrt{T}\left\{\left(\frac{x_t^\top x_t}{T}\right)^{-1} \frac{x_t^\top}{T}\left(\epsilon_t + x_t\beta_1 + \boldsymbol{x}_{-1,t}^\top \underbrace{\boldsymbol{\beta}_{-1,T}}_{:=\gamma/\sqrt{T}}\right) - \beta_1\right\}, \\ &\sqrt{T}\left\{\underbrace{\left(\frac{x_t^\top x_t}{T}\right)^{-1} \frac{x_t^\top \epsilon_t}{T}}_{(I)} + \beta_1 + \underbrace{\left(\frac{x_t^\top x_t}{T}\right)^{-1} \frac{x_t \boldsymbol{x}_{-1,t}}{T}\frac{\gamma}{\sqrt{T}}}_{(II)} - \beta_1\right\}. \end{aligned} \tag{1.14}$$

The term (I) is Gaussian by assumption on ϵ_t and β_1's cancel out. However, as $\mathbb{E}(x_t \boldsymbol{x}_{-1,t}) \neq 0$ by assumption, the term (II) produces a non-vanishing bias in expectation:

$$\begin{aligned} &\sqrt{T}\left\{\left(\frac{x_t^\top x_t}{T}\right)^{-1} \frac{x_t^\top \epsilon_t}{T} + \left(\frac{x_t^\top x_t}{T}\right)^{-1} \frac{x_t \boldsymbol{x}_{-1,t}}{T}\frac{\gamma}{\sqrt{T}}\right\} \\ &\xrightarrow{d} \mathcal{N}\left(\mathbb{E}(x_t^\top x_t)^{-1}\mathbb{E}(x_t\boldsymbol{x}_{-1,t})\gamma, \sigma^2\right). \end{aligned} \tag{1.15}$$

This bias hinders the coverage of any post-selection confidence interval. Therefore special techniques are needed to obtain uniform convergence to limit distributions. These techniques are referred to as "honest inference" and include: simultaneous inference across models (Berk et al., 2013), inference conditional on selected models (Lee et al., 2016), debiasing (or desparsifying) the lasso estimates (Van de Geer et al., 2014; Zhang and Zhang, 2014) and post-double-selection (PDS) techniques

(Belloni, Chernozhukov, and Hansen, 2014b).

1.1.2.1 The Post-Double Selection

Central to this thesis is the PDS method coined by Belloni, Chernozhukov, and Hansen (2014b). The method and the intuition of why this solves the pointwise convergence issue of post-selection estimators is now presented. Consider again as reference model (1.8) and again assume for simplicity $\boldsymbol{x}_{-1,t}$ has only one column. Consider $x_{1,t}$ the treatment variable i.e., the variable of interest for the inference. Then the following three steps, similarly to the famous Frisch-Waugh-Lovell theorem, are in order:

i) Step 1: Lasso of y_t on $\boldsymbol{x}_{-1,t}$:

$$\hat{\boldsymbol{\beta}}_{-1} = \arg\min_{\beta_{-1}} \left\| y_t - \boldsymbol{x}_{-1,t}^\top \boldsymbol{\beta}_{-1} \right\|_2^2 + \lambda \|\boldsymbol{\beta}_{-1}\|_1; \qquad \text{obtain} \quad \hat{S}_{\beta_{-1}}^{(I)}(\lambda).$$

ii) Step 2: Lasso of $x_{1,t}$ on $\boldsymbol{x}_{-1,t}$:

$$\hat{\boldsymbol{\beta}}_{-1} = \arg\min_{\beta_{-1}} \left\| x_{1,t} - \boldsymbol{x}_{-1,t}^\top \boldsymbol{\beta}_{-1} \right\|_2^2 + \lambda \|\boldsymbol{\beta}_{-1}\|_1; \qquad \text{obtain} \quad \hat{S}_{\beta_{-1}}^{(II)}(\lambda).$$

iii) Step 3: Least Squares of y_t on $x_{1,t}$ and $\boldsymbol{x}_{-1,t,\left(\hat{S}_{\beta_{-1}}^{(I)} \cup \hat{S}_{\beta_{-1}}^{(II)}\right)}$:

$$\hat{\beta}_1 = \arg\min_{\beta_1} \left\| y_t - \beta_1 x_{1,t} - \boldsymbol{x}_{-1,t,\left(\hat{S}_{\beta_{-1}}^{(I)} \cup \hat{S}_{\beta_{-1}}^{(II)}\right)}^\top \boldsymbol{\beta}_{-1} \right\|_2^2,$$

where $\hat{S}_{\beta_{-1}}^{(I)}(\lambda)$, $\hat{S}_{\beta_{-1}}^{(II)}(\lambda)$ are the estimated supports at Step 1 and 2 and $\left(\hat{S}_{\beta_{-1}}^{(I)} \cup \hat{S}_{\beta_{-1}}^{(II)}\right)$ indicates the union of the selected coefficients at Step 1 and 2. The intuition of this method is as follows. For simplicity assume the error term ϵ_t is homoskedastic and uncorrelated. Consider the covariance matrix

$$\mathbb{E}\begin{bmatrix} x_{1,t}^\top x_{1,t} & x_{1,t}\boldsymbol{x}_{-1,t} \\ x_{1,t}\boldsymbol{x}_{-1,t} & \boldsymbol{x}_{-1,t}^\top \boldsymbol{x}_{-1,t} \end{bmatrix} = \begin{bmatrix} \sigma^2_{x_{1,t}} & \sigma_{x_{1,t}}\sigma_{\boldsymbol{x}_{-1,t}}\rho \\ \sigma_{x_{1,t}}\sigma_{\boldsymbol{x}_{-1,t}}\rho & \sigma^2_{\boldsymbol{x}_{-1,t}} \end{bmatrix}, \tag{1.16}$$

where ρ is the correlation between $x_{1,t}$ and $\boldsymbol{x}_{-1,t}$ and assume $\rho \neq 0$. Now consider the regression of y_t on only $x_{1,t}$ i.e., $y_t = \beta_1 x_{1,t} + \tilde{\epsilon}_t$ where $\tilde{\epsilon}_t \equiv \boldsymbol{x}_{-1,t}^\top \boldsymbol{\beta}_{-1} + \epsilon_t$, then $\hat{\beta}_1^{\text{OLS}}$ will not be a consistent estimator for β_1 since

$$\begin{aligned}\hat{\beta}^{\text{OLS}} &= (x_{1,t}^\top x_{1,t})^{-1} x_{1,t}^\top (\beta_1 x_{1,t} + \tilde{\epsilon}_t) = \\ &= \beta_1 + (x_{1,t}^\top x_{1,t})^{-1} x_{1,t}^\top \tilde{\epsilon}_t = \beta_1 + (x_{1,t}^\top x_{1,t})^{-1} x_{1,t}^\top (\boldsymbol{x}_{-1,t}^\top \boldsymbol{\beta}_{-1} + \epsilon_t) \\ &= \beta_1 + (x_{1,t}^\top x_{1,t})^{-1} x_{1,t}\boldsymbol{x}_{-1}\boldsymbol{\beta}_{-1} + (x_{1,t}^\top x_{1,t})^{-1} x_{1,t}^\top \epsilon_t \\ &\xrightarrow{p} \beta_1 + \boldsymbol{\beta}_{-1}\frac{\rho\sigma_{x_{1,t}}}{\sigma_{\boldsymbol{x}_{-1,t}}}.\end{aligned}$$

where $\xrightarrow{p}$ indicates convergence in probability. Analogously, when considering the least squares of y_t on only $\boldsymbol{x}_{-1,t}$ will yields $\hat{\boldsymbol{\beta}}_{-1}^{OLS} \xrightarrow{p} \boldsymbol{\beta}_{-1} + \beta_1 \frac{\rho\sigma_{\boldsymbol{x}_{-1,t}}}{\sigma_{x_{1,t}}}$. Finally, taking the least squares of $x_{1,t}$ on $\boldsymbol{x}_{-1,t}$ returns

$$(\boldsymbol{x}_{-1,t}^\top \boldsymbol{x}_{-1,t})^{-1} \boldsymbol{x}_{-1,t}^\top x_{1,t} \xrightarrow{p} \frac{\rho\sigma_{x_{1,t}}\sigma_{\boldsymbol{x}_{-1,t}}}{\sigma^2_{\boldsymbol{x}_{-1,t}}}.$$

If $\boldsymbol{\beta}_{-1} + \beta_1 \frac{\rho\sigma_{\boldsymbol{x}_{-1,t}}}{\sigma_{x_{1,t}}}$ is large, then the lasso at Step 1 will select $\boldsymbol{x}_{-1,t}$. If $\frac{\rho\sigma_{x_{1,t}}}{\sigma_{\boldsymbol{x}_{-1,t}}}$ is also large, lasso will also select $\boldsymbol{x}_{-1,t}$ at Step 2. Conversely, $\frac{\rho\sigma_{x_{1,t}}}{\sigma_{\boldsymbol{x}_{-1,t}}}$ can only be small if ρ is small; hence $\boldsymbol{\beta}_{-1} + \beta_1 \frac{\rho\sigma_{\boldsymbol{x}_{-1,t}}}{\sigma_{x_{1,t}}}$ can only be small when $\boldsymbol{\beta}_{-1}$ is small. Therefore, only when ρ and $\boldsymbol{\beta}_{-1}$ are both small then lasso will not select $\boldsymbol{x}_{-1,t}$ in either Step 1 or Step 2 thus leaving Step 3 with no $\boldsymbol{x}_{-1,t}$. The key point is that if both ρ and $\boldsymbol{\beta}_{-1}$ are small, then

$$\hat{\beta}^{OLS} \xrightarrow{p} \beta_1 + \boldsymbol{\beta}_{-1}\frac{\rho\sigma_{x_{1,t}}}{\sigma_{\boldsymbol{x}_{-1,t}}} \approx \beta_1, \tag{1.17}$$

since the compounding factor $\boldsymbol{\beta}_{-1}\frac{\rho\sigma_{x_{1,t}}}{\sigma_{\boldsymbol{x}_{-1,t}}}$ is negligible and hence consistency is re-estabilished. In other words, considering equation (1.15), an unbiased result is obtained if $\sqrt{T}\left(\frac{x_t^\top x_t}{T}\right)^{-1}\frac{x_t^\top \boldsymbol{x}_{-1,t}}{T}\xrightarrow{p} 0$. Therefore, one can rewrite $x_{1,t} = \rho\frac{\sigma_{x_{1,t}}}{\sigma_{\boldsymbol{x}_{-1,t}}}\boldsymbol{x}_{-1,t}+\zeta_t$ with $\mathbb{E}(\zeta_t\boldsymbol{x}_{-1,t}) = 0$ and $\mathbb{E}\zeta^2 = 1-\rho^2$, and obtain

$$\begin{aligned}
&\left(T^{-1}\boldsymbol{x}_{-1,t}^\top\boldsymbol{x}_{-1,t}\right)^{-1}T^{-1/2}\left(\rho\frac{\sigma_{x_{1,t}}}{\sigma_{\boldsymbol{x}_{-1,t}}}\boldsymbol{x}_{-1,t}+\zeta\right)\boldsymbol{x}_{-1,t} = \\
&= \sqrt{T}\rho\frac{\sigma_{x_{1,t}}}{\sigma_{\boldsymbol{x}_{-1,t}}} + \left(T^{-1}\boldsymbol{x}_{-1,t}^\top\boldsymbol{x}_{-1,t}\right)^{-1}T^{-1/2}\boldsymbol{x}_{-1,t}^\top\zeta_t, \\
&= \sqrt{T}\rho\frac{\sigma_{x_{t,1}}}{\sigma_{\boldsymbol{x}_{-1,t}}} + (1-\rho^2)\frac{\sigma_{x_{1,t}}^2}{\sigma_{\boldsymbol{x}_{-1,t}}^2}\mathcal{N}(0,1) + o_p(1).
\end{aligned}$$

Define $\delta_T \to 0$ as $T\to\infty$ such that $\delta_T\sqrt{T}\to\infty$ but $\delta_T T^{1/4}\to 0$. Then, increasingly small parameters in T can be defined as $\rho = c_{x_{1,t}\boldsymbol{x}_{-1,t}}\delta_T$ and $\beta_T = c_{\boldsymbol{x}_{-1,t}}\delta_T$ for $c_{x_{1,t}\boldsymbol{x}_{-1,t}}, c_{\boldsymbol{x}_{-1,t}}$ some positive constants, such that

$$\begin{aligned}
&\beta_T\left(T^{-1}\boldsymbol{x}_{-1,t}^\top\boldsymbol{x}_{-1,t}\right)^{-1}T^{-1/2}\left(\rho\frac{\sigma_{x_{1,t}}}{\sigma_{\boldsymbol{x}_{-1,t}}}\boldsymbol{x}_{-1,t}+\zeta_t\right)\boldsymbol{x}_{-1,t} = \\
&= \sqrt{T}\delta_T^2 c_{x_{1,t}\boldsymbol{x}_{-1,t}}c_{\boldsymbol{x}_{-1,t}}\frac{\sigma_{x_{1,t}}}{\sigma_{\boldsymbol{x}_{-1,t}}} + c_{\boldsymbol{x}_{-1,t}}\delta_T(1-\rho^2)\frac{\sigma_{x_{1,t}}^2}{\sigma_{\boldsymbol{x}_{-1,t}}^2}\mathcal{N}(0,1),
\end{aligned} \tag{1.18}$$

and since $\sqrt{T}\delta_T^2\to 0$ and $\delta_T\to 0$ both terms vanish asymptotically, thus letting $\hat{\beta}_1^{\text{OLS}}\xrightarrow{p}\beta_1$. The double selection step via the lasso, guarantees that omitted variable bias is substantially diminished and the errors of the final model are close enough to be orthogonal with respect to the treatment variable. This rather straightforward result allows for uniform asymptotic validity for a test of hypothesis in a high-dimensional regression model.

1.2 Granger Causality

An important task investigated in this thesis is: trying to learn cause and effect relationships among time series in a high-dimensional modeling framework. Causality had been considered in the field starting around 1950 (see e.g. Weiner, 1956). However, it was Clive Granger's[9] contributions to the the study of causality and causal relationships in economics to pave the way to modern empirical causality analysis and testing. Granger (1969) Econometrica paper is the cornerstone of these fields as the simple definitions it contains have formed the basis for almost all the research in the area in the last 50 years and will likely do so for many more years to come. Granger employs spectral methods as well as simple bivariate time series models to formalize and illustrate the notion of causality. In his own words: "[...] Y_t causes X_t if we are able to better predict X_t using all available information than if the information apart from Y_t had been used"(Granger, 1969, p.428). Several research directions stemmed from this contribution to the literature: many forecasting works have used Granger causality tests as a basic tool for model specification and many economic theories like e.g., the relationship between money and income (see Sims, 1972) have been evaluated using Granger causality tests.

Later, Granger (1980) publishes "Testing for causality: a personal viewpoint" in the Journal of Economic Dynamics and Control. An elucidating discussion of the philosophical notion of causality and the roots of its initial interest is provided along with a probabilistic (axiomatic) formalization of the causality concept. The formal probabilistic interpretation of causality is derived in terms of distribution functions conditioned to an information set, thus leading to causality tests based on conditional expectation and variance.

Granger Causality captures predictability given a particular infor-

[9] Granger was awarded, together with Robert Engle, the Nobel Memorial Prize in Economic Sciences in 2003, in recognition of his contributions to cointegration analysis.

mation set Ω_t at time t. If the addition of the variable X_t to Ω_t alters the conditional distribution of another variable Y_t and both X_t and Ω_t are observed prior to Y_t, then X_t improves predictability of Y_t. Hence, we say X_t Granger causes (or is Granger causal for) Y_t.

Granger (1980) envisioned this information set Ω_t as "all the knowledge in the universe available at that time" (Granger, 1980, p.330). This is of course difficult to operationalize and poses some troubles. In fact, as observed in Eichler (2013), this probabilistic concept of causality exploits temporal precedence, namely the fact that causes must precede their effects in time. However, temporal precedence alone is not a sufficient condition for establishing cause–effect relationships, and the omission of relevant variables (cf. omitted variables bias), can lead to so-called spurious causalities. In other words, conditioning on an information set containing (all) the relevant variables is paramount to avoid confusing causal discoveries with mere predictability results i.e., causal results that do not hold anymore as an additional variable is added to Ω_t. Thus, the definition must be modified to become operational. To do so, one needs to substitute to the information set Ω_t, the set of all the information up until time t for the available data. For this operationalized version of causality Granger himself used the term "X_t is a prima facie cause of Y_t' to underline the fact that a cause in the sense of Granger causality must be considered only as a potential cause.

The high-dimensional setting under which this whole thesis is based on, allows to approach the original universal concept of causality as envisioned by Granger, thus rendering the operationalized version more robust. With high-dimensional models one is able to condition a relation between X_t and Y_t to a very large information set Ω_t. The curse of dimensionality as defined in Section 1.1.1 will constrain at giving up variables within the information set, unsuited for explaining the relationship among X_t and Y_t. However, the post-double selection algorithm outlined in Section 1.1.2 guarantees, within the possibilities of the data set available, that the information set selected is free of omitted variable bias.

1.3 High-Dimensional Time Series Models: Contribution of this Thesis

The present thesis is organized as follows.

Chapter 2 extends the post-double selection method as discussed in Section 1.1.2 to high-dimensional stationary time series models, specifically vector autoregressions (VARs). A Lagrange-Multiplier (LM) test is developed to test for (blocks) Granger causality in high-dimensional VARs. Through an extensive simulation study the test is proved to work very well in terms of both size and statistical power in finite samples. Many different ways of carefully tuning the penalty parameter λ are compared: information criteria, time series cross-validation and plug-in choices. The test is not confined to bivariate relations but accomodates blocks-Granger causality, meaning that a subset of variables can be tested to be Granger-causal for another set. Under a series of assumptions, the post-double selection estimator is proved to be asymptotically Gaussian and the relative LM test standard χ^2 distributed. The novel testing procedure is employed within the framework of a high-dimensional heterogeneous VAR (see Corsi, 2009) to build a contagion network of volatility spillovers for 30 large capital stocks. The proposed method is compared with standard bivariate Granger-causality and full system VAR Granger-causality tests and clusters of volatility contagion are derived via the edge betweenness algorithm. Comparisons are provided with both a large sample in which the full-system VAR provides a useful benchmark and a smaller sample. By increasing the information set through considering a high-dimensional VAR model in the estimation, one is able to obtain more realistic effects than in the low-dimensional models. Furthermore, even when the sample size is not large enough to use standard full-system VAR techniques, the proposed method remains reliable and delivers accurate results.

Chapter 3 builds on the post-double selection LM test for Granger causality in high-dimensional VARs developed in Chapter 2. The setting is extended to unit-root non-stationary time series. As usual asymptotic theory is not applicable to hypothesis testing in levels VARs if the variables are integrated or cointegrated, a lag-augmentation is employed similarly to Toda and Yamamoto (1995). While the original idea of Toda and Yamamoto (1995) was conceived for low-dimensional settings, in Chapter 2 this is extended to the high-dimensional case. Algebra is derived to show that augmenting the lag-length only to the variable of interest for the Granger causality test i.e., the Granger-causing and Granger-causal, as opposed to all the variables in the system, is sufficient to obtain asymptotic normality of the post-selection least squares estimator. Simulations show how this result is able to bypass the loss of statistical power produced by the lag-length over-specification as long as causality is tested on sufficiently small blocks. The set of assumptions needed for the post-double selection procedure to hold in the non-stationary framework is adapted and the LM test is again proved to be standard χ^2 distributed. Furthermore, a data-driven upper bound to select the lag-length in a high-dimensional VAR is proposed and its finite sample performances assessed. The test is used on the popular macroeconomics data set FRED-MD (see McCracken and Ng, 2016) to investigate the main macroeconomic drivers of inflation. The proposed method is able to uncover important macroeconomic connections which would be lost if differences would be taken to transform the time series to stationary.

Chapter 4 uses the designed post-double-selection LM test for unit root non-stationary time series developed in Chapter 3 to investigate causality in high-dimensional climate systems. The new method helps in disentangling and interpreting the complex causal chains linking greenhouse gas radiative forcings and global temperatures. Allowing for large-dimensionality opens up to opportunities of conditioning the causal relationship between greenhouse gases and temperature to several natural and anthropogenic variables. The use of a VAR in levels is particularly adapted for climate time series which are known to contain stochastic trends and yielding long memory. Climate change is discussed and in order to contribute to its attribution, Granger causality networks are built among the climate series considered via the post-double-selection LM test. Yearly data are collected from 1850 until 2019 on climate variables such as solar activity, stratospheric and tropospheric aerosols and surface albedo, ocean heat content, El Niño–Southern Oscillation index, global temperature anomalies and greenhouse gas concentration. GDP is also added as extra conditioned variable within the information set. We carry out the analysis both with greenhouse gas considered as a single aggregated series and also dividing it into the three main gases, namely CO_2, CH_4, N_2O. Direct and indirect causal paths are discussed as well as cycles, clusters and feedbacks effects. A sensitivity analysis on unit roots and lag-length show how avoiding taking differences of the original series is beneficial for the causal findings and considering larger lag-lengths is helpful for climate systems to uncover causal relations otherwise masked.

Chapter 5 reconciles sparse and dense techniques within the framework of a dynamic factor model. A two steps procedure is outlined in order to estimate the model and produce forecasts. The first step estimates the factor via standard principal components argument while the second step uses the estimated idiosyncratic components within a sparse VAR which is estimated by penalized regression techniques such as the adaptive lasso. Intuitively, this approach is beneficial since it allows to disentangle in the system covariance matrix, the dependence among its diverging eigenvalues, namely the factors, with the dependence among the bounded ones i.e., the idiosyncratic components. Cross-sectional and time dependence in the idiosyncratic term are allowed and this is assumed to follow a high-dimensional VAR model. Consistent estimation of both idiosyncratic components and the factors is shown as both the cross-sectional and time dimensions grow large. The work is complemented with a novel joint information criteria which combines the Bai and Ng (2002) approach to select the number of factors with an extra penalty which allows for simultaneous lag-length estimation. The forecasting performances of the proposed procedure as well as the proposed information criteria are assessed via simulations.

Concluding remarks are drawn in Chapter 6.

2

Granger Causality Testing in High-Dimensional VARs: a Post-Double-Selection Procedure[1]

[1] This chapter is based on a joint work with Alain Hecq and Stephan Smeekes from Maastricht University and it is forthcoming in the Journal of Financial Econometrics.

Abstract

In this chapter we develop an LM test for Granger causality in high-dimensional VAR models based on penalized least squares estimations. To obtain a test retaining the appropriate size after the variable selection done by the lasso, we propose a post-double-selection procedure to partial out effects of nuisance variables and establish its uniform asymptotic validity. We conduct an extensive set of Monte-Carlo simulations that show our tests perform well under different data generating processes, even without sparsity. We apply our testing procedure to find networks of volatility spillovers and we find evidence that causal relationships become clearer in high-dimensional compared to standard low-dimensional VARs.

2.1 Introduction

Economics, statistics and finance have seen a rapid increase of applications involving time series in high-dimensional systems. Central to many of these applications is the vector autoregressive (VAR) model that allows for a flexible modelling of dynamic interactions between multiple time series. In this chapter we develop a simple method to test for Granger causality in high-dimensional VARs (HD-VARs) with potentially many variables.

Many financial applications consider Granger causality analysis, especially for constructing high-dimensional networks. Networks of financial firms' intedependencies are investigated in Basu, Shojaie, et al. (2015), Gao et al. (2017), Demirer et al. (2018) and Barigozzi and Brownlees (2019). Similarly, spillovers and contagion among stock returns are investigated in networks using Granger causality analysis in Lin and Michailidis (2017), Vyrost et al. (2015) and Corsi et al. (2018).

Most of the econometric literature has traditionally been focused on allowing for high dimensionality in VARs through the use of factor models (see e.g. Bernanke et al., 2005; Chudik and Pesaran, 2016) or Bayesian methods (Bańbura et al., 2010). For instance Billio, Getmansky, et al. (2012) develops measures of connectedness to assess systemic risk propagation among institutions in the financial system using principal component analysis and Granger causality networks. Recent years have seen an increase in *regularized*, or penalized, estimation of *sparse* VARs based on popular methods from statistics such as the lasso (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005), which impose sparsity by setting a (data-driven) selection of the coefficients to zero.

Compared to factor models, such sparsity-seeking methods have often an advantage of interpretability, as in many economic applications, it appears natural to believe that the most important dynamic interactions among a large set of variables can be adequately captured by a relatively small – but unknown – number of 'key' variables. As such, the use of these methods for estimating HD-VAR models has also increased significantly in recent years, see e.g. Nicholson, Matteson, et al. (2017), Basu, Li, et al. (2019), Billio, Casarin, et al. (2019), Wilms and Croux (2018), Korobilis and Pettenuzzo (2019)).

Regularized estimation theory for high-dimensional time series and VAR models is now well established, see among others Song and Bickel (2011), Basu and Michailidis (2015a), Kock and Callot (2015), Davis et al. (2016), Medeiros and Mendes (2016a), Audrino and Camponovo (2018) and Masini et al. (2019) and Wong et al. (2020); Kock, Medeiros, et al. (2020) provide a recent review. However, performing inference on HD-VARs, such as testing for Granger causality, still remains a non-trivial matter. As is well known, performing inference after model selection (post-selection inference) is complicated as the selection step invalidates 'standard' inference where the uncertainty regarding the selection is ignored (see Leeb and Pötscher, 2005). Complexities introduced by the temporal and cross-sectional dependencies in the VAR mean that most recently developed post-selection inference methods are not automatically applicable.

Most existing literature on Granger causality testing in HD-VARs therefore has so far not considered post-selection inferential procedures.
Wilms, Gelper, et al. (2016) propose a bootstrap Granger causality test in HD-VARs, but do not account for post-selection issues. Similarly, Skripnikov and Michailidis (2019) investigate the problem of jointly estimating multiple network Granger causal models in VARs with sparse transition matrices using lasso-type methods, but focus mostly on estimation rather than testing. Song and Taamouti (2019) focus on statistical procedures for testing indirect/spurious causality in high-dimensional scenarios, but consider factor models rather than regularized regression techniques. Lin and Michailidis (2017) consider high-dimensional multi-block VARs derived from a two-blocks recursive linear dynamical system and use a maximum likelihood (ML) estimator for Gaussian data. In order to obtain the ML estimates for the system transition matrices and the precision matrix, respectively the lasso and graphical lasso on the residuals are iterated until convergence. Krampe, Kreiss, et al. (2018) develops bootstrap techniques for sparse

VAR models combining a model-based bootstrap procedure and the de-sparsified lasso (see Van de Geer et al. (2014)) to perform inference on the autoregressive parameters. Chaudhry et al. (2017) look at debiased estimators as in Javanmard and Montanari (2014), for Gaussian and sub-Gaussian VAR processes with a focus on Granger-causality and control of the false discovery rate.

In this chapter we build on the post-double-selection approach proposed by Belloni, Chernozhukov, and Hansen (2014b), to develop a valid post-selection test of Granger causality in HD-VARs. The finite-sample performance depends heavily on the exact implementation of the method. In particular, the tuning parameter selection in the penalized estimation is crucial. We therefore perform an extensive simulation study to investigate the finite-sample performance of the different ways to set up the test in order to be able to give some practical recommendations. In addition, we investigate the construction of networks of realized volatilities using a sample of 30 financial stocks modeled as a vector heterogeneous VAR (Corsi, 2009). We are able to demonstrate how our approach allows for obtaining much sharper conclusions than standard low-dimensional VAR techniques.

The remainder of the chapter is as follows: Section 2.2 introduces the high-dimensional VAR model and Granger causality tests. In Section 2.3 we propose our estimation and inferential framework. Section 2.4 establishes the asymptotic properties of our method and discusses the assumptions required for the theory to hold. Section 2.5 reports the results of the Monte Carlo simulations. We apply our method in Section 2.6 to construct volatility spillover networks. Section 2.7 concludes. Proofs and supplemental results can be found in the appendix.

A few words on notation. For any n-dimensional vector $\boldsymbol{x}$, we let $\|\boldsymbol{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ denote the ℓ_p-norm. For any index set $S \subseteq \{1, \ldots, n\}$, let $\boldsymbol{x}_S$ denote the sub-vector of $\boldsymbol{x}_t$ containing only those elements x_i such that $i \in S$. $|S|$ denotes the cardinality of the set S. We use $\xrightarrow{p}$ and $\xrightarrow{d}$ to denote convergence in probability and distribution, respectively.

2.2 High-dimensional Granger causality tests

Loosely speaking, the notion of Granger causality captures predictability given a particular information set (Granger, 1969; Granger, 1980). If the addition of variable X to the given information set Ω alters the conditional distribution of another variable Y, and both X and Ω are observed prior to Y, then X improves predictability of Y, and is said to *Granger cause* Y with respect to Ω. Granger (1969) originally envisioned the information set Ω "be all the information in the universe" (p. 428), which is of course not a workable concept. Yet clearly the choice of information set has a major effect on the interpretation of the finding of (non-)Granger causality, as discussed in Granger (1980). In particular, spurious Granger causality from X to Y may be found when both X and Y are Granger caused by Z, but Z is omitted from Ω. As such, one might want to include as many potentially relevant variables in the information set as possible in order to avoid finding spurious causality due to omitted variables, thereby moving as much as possible towards the universal information set envisioned by Granger. However, conditioning on so many variables leads to obvious problems of high-dimensionality rendering many standard statistical techniques invalid.

In this chapter we focus on testing Granger causality in mean using linear models, in which setup the VAR model is the natural tool to investigate this problem. However, to enlarge the information set means estimating a VAR with an increasing number of variables. The number of parameters in a VAR increases quadratically with the number of time series included; an unrestricted VAR(p) has K^2p coefficients to be estimated, where K is the number of series and p is the lag-length. As the time series dimension T is typically fairly small for many economic applications, the data do not contain sufficient information to estimate the parameters and consequently standard least squares and maximum likelihood methods suffer from the curse of dimensionality, resulting in estimators with high variance that overfit the data.

2.2.1 Granger causality testing in VAR models

Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$ be a K-dimensional multiple time series process, where $\boldsymbol{y}_t = (y_{1,t}, \ldots, y_{K,t})'$ is generated by a VAR(p) process

$$\boldsymbol{y}_t = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{A}_p y_{t-p} + \boldsymbol{u}_t, \qquad t = p+1, \ldots, T \ , \tag{2.1}$$

where for notational simplicity we assume the variables have zero mean; if not they can be demeaned prior to the analysis, or equivalently a vector of intercepts is added. $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p$ are $K \times K$ parameter matrices and $\boldsymbol{u}_t$ is a martingale difference sequence (mds) of error terms. We consider weakly stationary VAR models, as formalized in Assumption 1 below.

Assumption 1. The VAR model in (2.1) satisfies:

(a) $\{\boldsymbol{u}_t\}_{t=1}^T$ is a weakly stationary mds with respect to $\mathcal{F}_t = \sigma(\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, \ldots)$ $\boldsymbol{u}_t$ such that $\mathbb{E}(\boldsymbol{u}_t | \mathcal{F}_{t-1}) = \mathbf{0}$ for all t and $\boldsymbol{\Sigma}_u = \mathbb{E}(\boldsymbol{u}_t \boldsymbol{u}_t')$ is positive definite.

(b) All roots of $\det(\boldsymbol{I}_K - \sum_{j=1}^p \boldsymbol{A}_j z^j)$ lie outside the unit disc, such that the lag polynomial is invertible.

In the VAR model (2.1) we are interested in testing whether variables in the set J Granger cause variables in the set I in mean, conditional on all the other variables, where $J, I \subset \{1, \ldots, K\}$ and $J \cap I = \emptyset$. Let $N_I = |I|$ and $N_J = |N_J|$ denote the number of variables in I and J respectively. We describe our procedure here in general form for testing blocks of variables. For any sets $S_1, S_2 \subseteq \{1, \ldots, K\}$ of variables define the best linear predictor in L_2-norm of $\boldsymbol{y}_{S_1,t}$ given $\boldsymbol{x}_{S_2,t-1}^{(p)} = (\boldsymbol{y}_{S_2,t-1}', \ldots, \boldsymbol{y}_{S_2,t-p}')'$ as $\mathcal{P}(\boldsymbol{y}_{S_1,t} | \boldsymbol{x}_{S_2,t-1}^{(p)}) = \boldsymbol{\Gamma}^* \boldsymbol{x}_{S_2,t-1}^{(p)}$, where $\boldsymbol{\Gamma}^* = \min_{\boldsymbol{\Gamma}} \mathbb{E}\left[\|\boldsymbol{y}_{S_1,t} - \boldsymbol{\Gamma} \boldsymbol{x}_{S_2,t-1}\|_2^2 \right]$. Then we say that $\boldsymbol{y}_{J,t}$ *does not Granger cause* $\boldsymbol{y}_{I,t}$ conditionally on $\boldsymbol{x}_{J^c,t}$ if

$$\mathcal{P}(\boldsymbol{y}_{I,t} | \boldsymbol{x}_{J^c,t}^{(p)}) = \mathcal{P}(\boldsymbol{y}_{I,t} | \boldsymbol{x}_{t-1}^{(p)}) \tag{2.2}$$

for any value of $\boldsymbol{x}_{J^c,t}$. In other words, conditional on $\boldsymbol{x}_{J^c,t}$, addition of the lags of $\boldsymbol{y}_{J,t}$ to the information set does not improve predictability of $\boldsymbol{y}_{I,t}$. Note that Granger (non-)causality as defined in (2.2) is a property of the population. In the VAR (2.1) this means that testing for Granger causality can be done via testing the joint significance of the blocks of coefficients in the matrices $\boldsymbol{A}_1, \dots, \boldsymbol{A}_p$ corresponding to the impact of variables J on I.

To illustrate, consider (2.1) with $p = 1$ lag, and assume without loss of generality that the variables in $\boldsymbol{y}_t$ are ordered such that $\boldsymbol{y}_t = \left(\boldsymbol{y}'_{I,t}, \boldsymbol{y}'_{J,t}, \boldsymbol{y}'_{-(I\cup J),t}\right)'$, where $-(I \cup J))$ refers to all variables not in J or I. Then we can write

$$\begin{bmatrix} \boldsymbol{y}_{I,t} \\ \boldsymbol{y}_{J,t} \\ \boldsymbol{y}_{-(I\cup J),t} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{I,I} & \boldsymbol{A}_{I,J} & \boldsymbol{A}_{I,-(I\cup J)} \\ \boldsymbol{A}_{J,I} & \boldsymbol{A}_{J,J} & \boldsymbol{A}_{J,-(I\cup J)} \\ \boldsymbol{A}_{I,-(I\cup J)} & \boldsymbol{A}_{-(I\cup J)} & \boldsymbol{A}_{-(I\cup J),-(I\cup J)} \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_{I,t-1} \\ \boldsymbol{y}_{J,t-1} \\ \boldsymbol{y}_{-(J\cup I),t-1} \end{bmatrix} + \boldsymbol{u}_t, \tag{2.3}$$

where $\boldsymbol{A}$ is partitioned conformably with the blocks in $\boldsymbol{y}_t$. In this case, the best linear predictors in (2.2) are given by

$$\mathcal{P}(\boldsymbol{y}_{I,t}|\boldsymbol{y}_{t-1}) = \boldsymbol{A}_{I,I}\boldsymbol{y}_{I,t-1} + \boldsymbol{A}_{I,J}\boldsymbol{y}_{J,t-1} + \boldsymbol{A}_{I,-(I\cup J)}\boldsymbol{y}_{-(I\cup J),t-1},$$
$$\mathcal{P}(\boldsymbol{y}_{I,t}|\boldsymbol{y}_{J^c,t-1}) = \boldsymbol{A}^*_I\boldsymbol{y}_{J^c,t-1}, \text{ where } \boldsymbol{A}^*_I = \min_{\boldsymbol{A}_I} \mathbb{E}\left[\|\boldsymbol{y}_{I,t} - \boldsymbol{A}_I\boldsymbol{y}_{J^c,t-1}\|_2^2\right].$$

For any arbitrary value of $\boldsymbol{y}_{t-1}$, these can only coincide if $\boldsymbol{A}_{I,J} = \boldsymbol{0}$. Hence, the null hypothesis of no Granger causality from J to I in the VAR(1) model can be formulated in terms of $\boldsymbol{A}_{I,J} = \boldsymbol{0}$. This is easily extended to $p > 1$ by simply testing if the (I, J)-block of all p lag matrices is equal to zero.

In the remainder of the chapter, we will be working with a stacked representation of (2.1) for the variables in I. Specifically, let $\boldsymbol{Y} = (\boldsymbol{y}_{p+1}, \dots, \boldsymbol{y}_T)'$ and let $\boldsymbol{y}_I = \operatorname{vec}(\boldsymbol{Y}_I)$ denote the $N_I \times 1$ stacked vector containing all observations corresponding to the variables in I. Similarly, let $\boldsymbol{u}_I = \operatorname{vec}(\boldsymbol{U}_I)$, where $\boldsymbol{U} = (\boldsymbol{u}_{p+1}, \dots, \boldsymbol{u}_T)'$. Let $\boldsymbol{X} = \left(\boldsymbol{x}_p^{(p)}, \dots, \boldsymbol{x}_{T-1}^{(p)}\right)'$ and $\boldsymbol{X}^{\otimes} = \boldsymbol{I}_{N_I} \otimes \boldsymbol{X}$, while defining the stacked parameter vector $\boldsymbol{\beta} = \text{vec}((\boldsymbol{A}_1, \dots, \boldsymbol{A}_p)')$. Then we can write

$$\boldsymbol{y}_I = \boldsymbol{X}^{\otimes}\boldsymbol{\beta} + \boldsymbol{u}_I = \boldsymbol{X}^{\otimes}_{GC}\boldsymbol{\beta}_{GC} + \boldsymbol{X}^{\otimes}_{-GC}\boldsymbol{\beta}_{-GC} + \boldsymbol{u}_I, \tag{2.4}$$

where $\boldsymbol{X}^{\otimes}_{GC} = \boldsymbol{I}_{N_I} \otimes \boldsymbol{X}_{GC}$, and $\boldsymbol{X}_{GC} = \left(\boldsymbol{x}^{(p)}_{J,p}, \dots, \boldsymbol{x}^{(p)}_{J,T-1}\right)'$ contains those columns of $\boldsymbol{X}$ corresponding to the potentially Granger causing variables in J; $\boldsymbol{X}_{-GC}$ and $\boldsymbol{X}^{\otimes}_{-GC}$ are then defined similarly but containing the remaining variables.[2] Testing for no Granger causality is then equivalent to testing $H_0 : \boldsymbol{\beta}_{GC} = \boldsymbol{0}$ against $H_1 : \boldsymbol{\beta}_{GC} \neq \boldsymbol{0}$.

Define $N_J = |J|$ and $N_I = |I|$. Note that $\boldsymbol{\beta}_{-GC}$ has $(K - N_J) \times N_I \times p$ elements, which we assume large through having a large number of variables K. On the other hand, throughout the chapter we assume that N_J, N_I and p are small, or more precisely, fixed when sample size increases to infinity. As $\boldsymbol{\beta}_{GC}$ has $N_{GC} = N_J \times N_I \times p$ elements, these are also implied to be fixed. While theoretically it is possible to consider an increasing number of elements in $\boldsymbol{\beta}_{GC}$ (see Remark 2.6 for details), it would not be required for typical applications. J and I are under the researcher's control and in most applications it is natural to consider a small number of variables of interest; often both J and I will only consist of a single variable, as in our application.

For p it may appear more restrictive to assume it small. However, large p in univariate regressions or small systems often arise from neglected dynamics with omitted variables (Hecq et al., 2016). As our HD-VAR attempts to include many more variables than typical small systems, we hope to alleviate the omitted variable issue, and thereby also directly making smaller p much more realistic. Of course, p is generally unknown in practice. However, in many applications it is possible to give a reasonable (and small) upper bound on p, which is sufficient for our algorithm. If not, p has to be estimated. We discuss two ways in the next section.

[2] Note that if $I = \{i\}$ for one particular value of interest, then (2.4) simply corresponds to a single equation from the VAR in (2.1).

Remark 2.1. Our operational version of Granger causality only considers causality in mean. Additionally, one might argue that considering only linear models is a further restriction on the generality of the concept of Granger causality. However, in our high-dimensional approach linear models are less restrictive as would appear. First, the VAR does not have to be formulated for levels of variables of interest. In fact, in our application we formulate a VAR for (realized) variances, such that we are implicitly testing Granger causality in second moments rather than first moments. Second. the linear VAR model in many cases provides a good approximation to a general nonlinear process via a Wold-type representation argument, see e.g. Meyer and Kreiss (2015). Finally, non-linear transformations (such as powers) of the original variables can be added to (2.4), by which general functional forms can be approximated (even if one then strictly loses the VAR equivalence). While in small systems this is infeasible as it increases the dimensionality disproportionally, our high-dimensional approach can handle this without any conceptual issues. In fact, Belloni, Chernozhukov, and Hansen (2014a) explicitly motivate their high-dimensional linear approach as an approximation to a general function; their arguments apply here as well.

2.3 Inference after selection by the lasso

In this section we introduce our inferential procedure to the Granger causality tests in high-dimensional VARs. We first discuss the lasso, which we use in the initial stage to select relevant variables. Next we discuss how naive use of the lasso introduces post-selection problems for inference, and we propose our algorithm to remedy this.

2.3.1 The lasso estimator

As $\boldsymbol{\beta}$ is high-dimensional when Kp is large relative to T, least squares estimation is not appropriate, and a structure must be imposed on $\boldsymbol{\beta}$ to be able to estimate it consistently. We assume *sparsity* of $\boldsymbol{\beta}$; that is, we assume that $\boldsymbol{\beta}$ can accurately be approximated by a coefficient vector with a (significant) portion of the coefficients equal to zero.

The sparsity assumption validates the use of variable selection methods, thereby reducing the dimensionality of the system without having to sacrifice predictability. For a general n-dimensional vector of responses $\boldsymbol{y}$ and $n \times M$-dimensional matrix of covariates $\boldsymbol{X}$, the (weighted) lasso simultaneously performs variable selection and estimation of the parameters by solving

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{T} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{m=1}^{M} |w_m \beta_m| \right), \tag{2.5}$$

where λ is a non-negative tuning parameter determining the strength of the penalty, and $\{w_m\}_{m=1}^{M}$ are non-negative weights corresponding to the parameters in $\boldsymbol{\beta}$. For the standard lasso the weights are either equal to one, or equal to zero (if this parameter should not be penalized). The notation $\hat{\boldsymbol{\beta}}(\lambda)$ highlights that the solution to the minimization problem depends on λ, which has te be selected as well (see Section 2.3.3). When no confusion can arise, we simply write $\hat{\boldsymbol{\beta}}$.

One may also consider the *adaptive lasso* (Zou, 2006) with parameter-specific weights w_j in (2.5) based on an initial estimation of $\boldsymbol{\beta}$, which is able to delete more irrelevant variables. However, for our purpose such oracle properties are not very relevant; we wish to eliminate the effects of the other "nuisance" variables on the relation between the variables tested for Granger causality, but we do not need to identify which of these nuisance variables matter.

Theoretical properties of lasso estimation in stable VAR models have now been studied extensively. We here non-exhaustively mention some of the key results for our setting; see Kock, Medeiros, et al. (2020) for a thorough review. Kock and Callot (2015) derive oracle properties of the adaptive lasso for VAR models. Basu and Michailidis (2015a) establish restricted eigenvalue conditions for VAR models and show their sufficiency for estimation consistency. Medeiros and Mendes (2016a) relax the Gaussianity assumptions of these papers by considering conditionally heteroskedastic errors, and demonstrate that the adaptive lasso retains oracle properties in time series settings. Finally, Masini et al. (2019) derive bounds on estimation errors in approximately sparse VAR models under very general conditions, allowing for heavy tails and dependence in the error terms. In particular, they show that several commonly used volatility processes in financial research satisfy these assumptions, thereby formally establishing the suitability of the lasso for many financial applications of VAR models.

2.3.2 Post-Double-Selection Granger causality test

2.3.2.1 The need for post selection inference

One might be tempted to simply perform the (adaptive) lasso as in (2.5) on (2.4), setting $w_{GC} = 0$, and then testing whether $\boldsymbol{\beta}_{GC} = 0$, potentially after re-estimating the model by OLS on only the selected variables. However, this ignores the fact that the final, selected, model is random and a function of the data. The randomness contained in the selection step means the post-selection estimators do not converge uniformly to a normal distribution, as the potential omitted variable bias from omitting (weakly) relevant variables in the selection step is too large to maintain uniformly valid inference.

In a sequence of papers (see e.g. Leeb and Pötscher, 2005), Leeb and Pötscher address these issues, showing that distributions of post-selection estimators only converge point-wise but not uniformly in the parameter space to normal distributions. Therefore, “standard” asymptotics fail to deliver a proper approximation of finite-sample behavior due to the presence of small, hard to detect parameters, whose omitted variable bias is too large to ignore asymptotically. As such, post-selection based on oracle properties is only appropriate if one a priori rules out small parameters conditions (via *beta-min* conditions, see e.g. Geer, Bühlmann, et al., 2011) thus obtaining a sharp separation of non-zero from zero coefficients. This is typically far too strong to be reasonable in applications, and methods explicitly accounting for selection are required.

Several approaches to valid post-selection inference, also referred to as *honest inference*, have been developed in recent years based on various philosophies, such as simultaneous inference across models (Berk et al., 2013), inference conditional on selected models (Lee et al., 2016), or debiasing (desparsifying) the lasso estimates (Van de Geer et al., 2014; Zhang and Zhang, 2014). We focus on the double selection approach developed by Belloni, Chernozhukov and co-authors; see e.g. (Belloni, Chernozhukov, and Hansen, 2014a) for an overview. This approach is tailored for the lasso, easy to implement, and can be extended to dependent data.

Belloni, Chernozhukov, and Kato (2014) develop a *post-double-selection* approach to construct uniform inference for treatment effects in partially linear models with high-dimensional controls using the lasso. Two initial lasso estimations of both the outcome and the treatment variable on all the controls are performed, and a final post-selection least squares estimation is conducted of the outcome variable on the treatment variable and all the controls selected in *at least* one of the two steps. The double variable selection step substantially diminishes the omitted variable bias and ensures the errors of the final model are (close enough to) orthogonal with respect to the treatment. The authors proved uniform validity of the procedure under a wide range of DGPs, including heteroskedastic and non-Gaussian errors.

Chernozhukov, Härdle, et al. (2020) extend the analysis of estimation and inference for highly-dimensional systems in regressions, allowing for (weak) temporal and cross-sectional dependency. Regularization techniques for dimensionality reduction are applied iteratively in the system and the overall penalty is jointly chosen by a block multiplier bootstrap procedure. Oracle properties and bootstrap consistency of the test procedure are derived. Furthermore, simultaneous valid inference is obtained via algorithms employing least square or least absolute deviation after (double) lasso selection step(s). Although our approach is closely related to that of Chernozhukov, Härdle, et al. (2020), it differs in a number of ways. Our method is simpler and faster to implement as it does not rely on bootstrap methods. Also, Chernozhukov, Härdle, et al. (2020) focus on general systems of equations and general ways of performing inference, which is different from our specific focus on Granger causality and VAR models. Third, we consider a different set of assumptions to establish the validity of our method, where we specifically focus on the relevance of these assumptions for applications in financial econometrics.

2.3.2.2 High-dimensional Granger causality test

We here describe how to implement the post-double-selection procedure in a VAR context. Let $\boldsymbol{x}_{GC,j}$, $j = 1, \dots, N_X$, where $N_X = pN_J$, denote the j-th column of $\boldsymbol{X}_{GC}$ and consider the partial regressions:

$$\boldsymbol{y}_I = \boldsymbol{X}^{\otimes}_{-GC}\gamma_0 + \boldsymbol{e}_0, \tag{2.6}$$

$$\boldsymbol{x}_{GC,j} = \boldsymbol{X}_{-GC}\boldsymbol{\gamma}_j + \boldsymbol{e}_j, \qquad j = 1, \dots, N_X, \tag{2.7}$$

where $\boldsymbol{\gamma}_j$, $j = 0, \dots, N_X$, are the best linear prediction coefficients[3]

$$\gamma_0 = \arg\min_{\boldsymbol{\gamma}} \mathbb{E}\left\|\boldsymbol{y}_{I,t} - \boldsymbol{X}^{\otimes\prime}_{-GC,t-1}\gamma\right\|_2^2 = \left(\mathbb{E}\boldsymbol{X}^{\otimes}_{-GC,t-1}\boldsymbol{X}^{\otimes\prime}_{-GC,t-1}\right)^{-1}\mathbb{E}\boldsymbol{X}^{\otimes}_{-GC,t-1}\boldsymbol{y}_{i,t},$$

$$\gamma_j = \arg\min_{\boldsymbol{\gamma}} \mathbb{E}\left\|\boldsymbol{x}_{GC,j,t} - \boldsymbol{x}'_{-GC,t-1}\gamma\right\|_2^2 = \left(\mathbb{E}\boldsymbol{x}_{-GC,t-1}\boldsymbol{x}'_{-GC,t-1}\right)^{-1}\mathbb{E}\boldsymbol{x}_{-GC,t-1}\boldsymbol{x}_{GC,j,t},$$

for $j = 1, \dots, N_X$, where $\boldsymbol{X}^{\otimes}_{-GC,t-1} = \boldsymbol{I}_{N_I} \otimes \boldsymbol{x}_{-GC,t-1}$. As the errors $\boldsymbol{e}_0, \dots, \boldsymbol{e}_{N_X}$ are orthogonal to $\boldsymbol{X}_{-GC}$, partialling out the effects of these variables would allow for a valid test of Granger causality. Of course, (2.6) and (2.7) are still high-dimensional and cannot be estimated by least squares. However, we can select the relevant variables from lasso estimation of (2.6) and (2.7) and collect all these for the final estimation of $\boldsymbol{y}_I$ on $\boldsymbol{X}^{\otimes}_{GC}$ plus only those relevant variables.

[3] Note that Assumption 2(a) implies that $\left(\mathbb{E}\boldsymbol{x}_{-GC,t-1}\boldsymbol{x}_{-GC,t-1}\right)^{-1}$ and hence $\left(\mathbb{E}\boldsymbol{X}^{\otimes}_{-GC,t-1}\boldsymbol{X}^{\otimes\prime}_{-GC,t-1}\right)^{-1}$ exists.

Intuitively, this works because to cause omitted variable bias on the coefficients of $\boldsymbol{X}_{GC}$, a particular variable in $\boldsymbol{X}_{-GC}$ must have a nonzero coefficient in *both* (2.6) and one of the regressions in (2.7). If its coefficient is zero in (2.6), it has no effect on $\boldsymbol{y}_I$ and is therefore not wrongfully omitted. If it has a zero coefficient in all regressions in (2.7), it is not correlated with any variables of interest, and omitting it will not result in a bias. By including all variables that are selected in at least a single of these regressions, we essentially allow for "one free mistake" by the lasso in failing to select a relevant variable. That is, omitted variable bias will only occur if the lasso fails to select a relevant variable in both regressions simultaneously. As the probability of this occurring decreases quadratically, this is sufficient to be negligible asymptotically and allow for uniformly valid inference. We provide a formal justification in Section 2.4.

We now state the details of our algorithm which executes the post-double-section along the lines described above, and conclude this section with some remarks.

Remark 2.2. We perform the initial regressions in terms of $\boldsymbol{X}_{GC}$ amd $\boldsymbol{X}_{-GC}$ instead of $\boldsymbol{X}^{\otimes}_{GC}$ and $\boldsymbol{X}^{\otimes}_{-GC}$. The two are equivalent, as the Kronecker product essentially just copies the columns of $\boldsymbol{X}$ both in the dependent and explanatory variables. Running the initial regressions in terms of $\boldsymbol{X}^{\otimes}$ therefore essentially means running the same regression N_I times, which is pointless as the selected variables remain the same in terms of the columns of $\boldsymbol{X}_{-GC}$. We therefore perform the regressions just once for each column in $\boldsymbol{X}_{GC}$. The construction of $\hat{S}^{\otimes}_X$ ensures that for any selected column $\boldsymbol{x}_{-GC,m}$, we select every column of $\boldsymbol{X}^{\otimes}_{-GC}$ in which $\boldsymbol{x}_{-GC,m}$ appears.

Remark 2.3. The feasible generalized least squares (FGLS) estimation in Step [2] is needed when $N_I > 1$ to account for the correlation between equations of the VAR, and the fact we do not have the same selected regressors in each equation, as those coming from (2.6) differ. Note that if $N_I = 1$, FGLS estimation collapses to the familiar form of the

Algorithm 1 Post-double-selection Granger causality LM test (PDS-LM)

[1] Estimate the initial partial regressions in (2.6) and (2.7) by an appropriate sparsity-inducing estimator such as the (adaptive) lasso, and let $\hat{\gamma}_0, \ldots, \hat{\gamma}_{N_X}$ denote the resulting estimators. Let $\hat{S}_0 = \{m : |\hat{\gamma}_{m,0}| > 0,\ m = 1, \ldots, N\}$ and $\hat{S}_j = \{m : |\hat{\gamma}_{m,j}| > 0,\ m = 1, \ldots, N_X\}$ for $j = 1, \ldots, p$, denote the selected variables in each regression.

[2] Let $\hat{S}_X = \bigcup_{j=1}^{N_X} \hat{S}_j$ denote all variables selected in the regressions for the columns of $\boldsymbol{X}_{GC}$, and let $\hat{S}_X^{\otimes}$ map $\hat{S}_X$ back to $\boldsymbol{X}_{-GC}^{\otimes}$ be such that $\boldsymbol{X}^{\otimes}_{\hat{S}_X^{\otimes}} = \boldsymbol{I}_{N_I} \otimes \boldsymbol{X}_{\hat{S}_X}$. Collect all variables kept by the lasso in Step [1] in $\hat{S}^{\otimes} = \hat{S}_0 \cup \hat{S}_X^{\otimes}$. Obtain the residuals $\hat{\boldsymbol{\xi}} = \boldsymbol{y}_I - \boldsymbol{X}^{\otimes}_{\hat{S}^{\otimes}} \hat{\boldsymbol{\beta}}^{\dagger}$ by OLS estimation. Let $\hat{\boldsymbol{\Xi}}_I$ denote the $T \times N_I$-matrix formed from $\hat{\boldsymbol{\xi}}$ and construct $\hat{\boldsymbol{\Sigma}}_{u,I} = \hat{\boldsymbol{\Xi}}_I' \hat{\boldsymbol{\Xi}}_I / T$ and $\hat{\boldsymbol{\Sigma}}_{u,I}^{\otimes} = \hat{\boldsymbol{\Sigma}}_{u,I} \otimes \boldsymbol{I}_T$.

[3] Let $\boldsymbol{y}_{N_I}^{*} = \left(\hat{\boldsymbol{\Sigma}}_{u,I}^{\otimes}\right)^{-1/2} \boldsymbol{y}_{N_I}$ and $\boldsymbol{X}^{*\otimes} = \left(\hat{\boldsymbol{\Sigma}}_{u,I}^{\otimes}\right)^{-1/2} \boldsymbol{X}^{\otimes}$. Obtain the residuals $\hat{\boldsymbol{\xi}}^{*} = \boldsymbol{y}_I^{*} - \boldsymbol{X}^{*\otimes}_{\hat{S}^{\otimes}} \hat{\boldsymbol{\beta}}^{\dagger}_{FGLS}$, and regress $\hat{\boldsymbol{\xi}}^{*}$ onto the variables retained by the previous regularization steps plus the Granger causality variables, retaining the residuals $\hat{\boldsymbol{\nu}}^{*} = \hat{\boldsymbol{\xi}}^{*} - \boldsymbol{X}^{*\otimes}_{\hat{S} \cup GC} \hat{\boldsymbol{\beta}}^{*}_{FGLS}$. Then obtain the statistic $LM = (\hat{\boldsymbol{\xi}}^{*\prime} \hat{\boldsymbol{\xi}}^{*} - \hat{\boldsymbol{\nu}}^{*\prime} \hat{\boldsymbol{\nu}}^{*})$.

[4a] Reject H_0 if $LM > q_{\chi^2_{N_{GC}}}(1-\alpha)$, where $q_{\chi^2_{N_{GC}}}(1-\alpha)$ is the $1-\alpha$ quantile of the χ^2 distribution with N_{GC} degrees of freedom.

[4b] Reject H_0 if $\left(\frac{TN_I - \hat{s} - N_{GC}}{N_{GC}}\right)\left(\frac{LM}{TN_{GC} - LM}\right) > q_{F_{N_{GC}, N_I T - \hat{s} - N_{GC}}}(1-\alpha)$, where $\hat{s} = \left|\hat{S}^{\otimes}\right|$ and $q_{F_{N_{GC}, N_I T - \hat{s} - N_{GC}}}(1-\alpha)$ is the $1-\alpha$ quantile of the F distribution with N_{GC} and $N_I T - \hat{s} - N_{GC}$ degrees of freedom.

LM statistic. In that case one regresses $\hat{\boldsymbol{\xi}}$ by OLS onto the variables retained by the previous regularization steps plus the Granger causality variables, and retain the residuals $\hat{\boldsymbol{\nu}} = \hat{\boldsymbol{\xi}} - \boldsymbol{X}^{\otimes}_{\hat{S}\,\cup\,GC}\hat{\boldsymbol{\beta}}^{*}$, obtaining $R^2 = 1 - \hat{\boldsymbol{\nu}}'\hat{\boldsymbol{\nu}}/\hat{\boldsymbol{\xi}}'\hat{\boldsymbol{\xi}}$.

Remark 2.4. Our lasso estimation of an HD-VAR can be interpreted as a general, data-driven, approach to Granger causality testing which encompasses the theory-driven 'standard' approach in low-dimensional VARs. In particular, the lasso can be interpreted as imposing (approximate) sparsity over a high-dimensional information set, with the extent and location of the sparsity, or irrelevance, determined in a data-driven way. Conversely, testing Granger causality in a low-dimensional setting can then be interpreted as a priori assuming an extreme degree of sparsity over the same information set; in other words, it amounts to assuming that none of the additional series are relevant.

Remark 2.5. Given that we essentially have $N_{GC} = N_J \times N_I \times p$ steps of selection, it would be more appropriate to refer to our method as "post-N_{GC}-selection" approach. For expositional simplicity however we stick to the post-double-selection name, as this is the common name for such a procedure, and conveys the essence of our method equally well.

Remark 2.6. Although the lasso regressions can handle increasing N_J, N_I or p with any issues, inference becomes more complicated when N_{GC} increases with the sample size as the proposed LM statistic (or similarly a Wald test) will not have a limit distribution anymore. In such a case one could use recently developed Gaussian approximations of maxima of high-dimensional vectors (Chernozhukov, Chetverikov, et al., 2013; Zhang and Wu, 2017) to base a test statistic on $\max_{m=1,\dots,N_{GC}} \left|\hat{\beta}^{\text{PDS}}_{GC,m}\right|$, where $\hat{\boldsymbol{\beta}}^{\text{PDS}}_{GC}$ are the coefficients of $\boldsymbol{X}^{\otimes}_{GC}$ in a regression of $\boldsymbol{y}_I$ on $\boldsymbol{X}_{GC}$ and $\boldsymbol{X}^{\otimes}_{\hat{S}^{\otimes}}$ as in (2.8). However, the critical values of this test statistic have to be simulated, which complicates the testing. As we argued in Section 2.2 that a fixed N_{GC} is a reasonable assumption for typical Granger causality applications, we do not pursue this route.

Remark 2.7. In Step [1] we propose not to consider the GC variables in the first regularization and insert them back at Step [2]. Alternatively, the GC variable(s) can be left in the regression, such that, we regress on the full $\boldsymbol{X}^{\otimes}$ matrix. In this case there are then two further possibilities by either penalizing these variables or not. Simulations for these two alternatives have been carried out and in practice we do not find significant differences among the three in terms of size and power. The approach proposed in Step 1 delivers the best results in terms of size.

Remark 2.8. When the time series length is of same magnitude as the number of covariates, information criteria and time series cross-validation tend to break down and select too many covariates in order to perform a post-selection by OLS. To overcome this issue we propose to place a lower bound on the penalty to ensure that in each selection regression at most cTN_I variables are selected, for some $0 < c < 1$. In our simulation and empirical studies we set $c = 0.5$. Note that, as we have N_{GC} selection steps, the possibility remains that different variables are selected in each steps, making the number of variables in the union $\hat{s}$ still too large to perform the post-selection OLS, although this problem is likely to occur far less often. This can be addressed by ensuring that fewer than $N_I T/N_{GC} = T/N_X$ variables are selected in each selection step. We do not impose this stricter bound in general, as it will often be much too strict. Instead, we recommend to only address this issue if it arises in practice by an ad-hoc increase of the lower bound on the penalty.[4]

[4] Although it happens less often, the theoretical plug-in method for the tuning parameter occasionally also selects too many variables to make the post-OLS estimation infeasible. However, for this method no easy solution is available for bounding the penalty. One could increase the constant in the plug-in expression, thus strengthening the penalty, but this would be a rather ad-hoc adjustment. In particular, imposing the lower bound for the other methods only limits the allowed

Remark 2.9. Although our Granger causality test has a χ^2 distribution under the null hypothesis asymptotically, in smaller samples the test might still suffer from the usual small-sample approximation error. As such we propose a finite-sample correction to the test in Step [3b], which in our simulation studies improved the size of our test.

Remark 2.10. Instead of the (adaptive) lasso, other estimators can be used in Step [1] as long as they deliver a sparse coefficient vector. For instance, the elastic net of Zou and Hastie (2005) that adds an ℓ_2-penalty in addition to the ℓ_1-penalty of the lasso can be used. The additional penalty ensures that the elastic net is strictly convex, and as a consequence tends to select highly correlated variables as a group together, whereas the lasso would tend to select only one of these variables (Zou and Hastie, 2005). Given the typically strong correlations between many economic variables, this appears particularly useful for our context. However, we used the elastic net for both the simulations and the empirical application, and in both cases we found that the results are widely comparable to those of lasso. Therefore we chose to omit them from the chapter.

Remark 2.11. One can also perform a standard Wald test of Granger causality instead of the LM test, by regressing the variables of interest on $\boldsymbol{X}_{GC}^{\otimes}$ and $\boldsymbol{X}_{\hat{S}^{\otimes}}^{\otimes}$, and testing for the significance of the coefficients of $\boldsymbol{X}_{GC}^{\otimes}$. While asymptotically the LM and Wald tests behave equally, differences might arise in small samples. We investigated the Wald version of the test in simulations as well, with results reported in Appendix B, Table 2.3. In general, differences between the two methods are negligible. However, for the Wald test, occasionally we run into the problem described in Remark 2.8, where even with the imposed lower

range of the tuning parameter, forcing the minimization to choose another (local) minimum that can still be far away from the boundary and justified graphically. For the plug-in method it is however difficult to justify the right amount of the increase, as the tuning parameter will be fixed to that value, and thus the chosen increase is rather arbitrary.

bound on the penalty, too many variables are selected for performing a post-selection OLS. For this reason we prefer the LM version.

2.3.3 Tuning parameter selection

Appropriate selection of the lasso tuning parameter λ in (2.5) is crucial to achieve good performance. Many different data-driven methods exist giving wildly varying results. We provide a systematic comparison of several popular methods discussed in the literature in a simulation study. To the best of our knowledge, this is the first such comparison in the context of post-selection inference. We now introduce the methods considered in our study.

One option is to minimize an information criterion (IC) to determine an appropriate data-driven λ. Let $\hat{S}(\lambda) = \left\{m \in \{1, \ldots, Kp\} : \left|\hat{\beta}_m(\lambda)\right| > 0\right\}$ denote the set of active variables in the lasso solution for a given λ. For a generic response vector $\boldsymbol{y}$ and predictor matrix $\boldsymbol{X}$, the value λ^{IC} is found as

$$\lambda^{IC} = \arg\min_{\lambda} \ln\left(\frac{1}{T}\left\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda)\right\|_2^2\right) + \frac{C_T}{T}\left|\hat{S}(\lambda)\right|,$$

where C_T is the penalty specific to each criterion. We consider the *Akaike information criterion* (AIC) by Akaike (1974) with $C_T = 2$, the *Bayesian information criterion* (BIC) by Schwarz (1978) with $C_T = \ln(T)$, and the *Extended Bayesian information criterion* (EBIC) by Chen and Chen (2008) with $C_T = \ln(T) + 2\gamma\ln(Kp)$ with $\gamma = 0.5$ proposed by Chen and Chen (2012) who argue that BIC fails to select the correct variables when the number of parameters is larger than the sample size.

An alternative approach is to plug in estimates of theoretically optimal values (see e.g. Bickel, Ritov, et al., 2009; Belloni and Chernozhukov, 2013; Belloni, Chernozhukov, and Wang, 2011). The lasso requires that $\lambda \geq c\|\boldsymbol{X}'\boldsymbol{u}\|_\infty/T$ for some constant $c > 0$ with "high probability". The central limit theorem motivates a Gaussian approximation where one chooses $\lambda^{th} = \frac{2c\hat{\sigma}}{\sqrt{T}}\Phi^{-1}\left(1 - \frac{\alpha}{2N}\right)$ for a small $\alpha = o(1)$, where $\Phi^{-1}(\cdot)$ is the inverse of standard Gaussian cumulative distribution function and $\hat{\sigma}$ is an estimate the variance of $\boldsymbol{u}$. In this chapter we set $\alpha = 0.05/\ln(T)$ and $c = 0.5$, while we follow Belloni, Chen, et al. (2012) in the estimation of σ. Specifically, we obtain an initial (conservative) estimate by least squares estimation of $\boldsymbol{y}$ on the five most correlated regressors. This estimate is then updated iteratively, for details see Belloni, Chen, et al. (2012).

Perhaps the most popular way to choose the tuning parameter is cross-validation (CV), although CV is not always appropriate in the time series setup without modifications (Bergmeir et al., 2018). To estimate the tuning parameter with CV in a time series setup (TSCV) we use an expanding window out-of-sample forecasting scheme and minimize its squared forecasting error. The rolling window is set up with 80% of the sample for training and 20% for testing. Cross-validation is appealing since it does not require any plug-in estimates, however, as observed in Chetverikov et al. (2020) it typically yields small values of λ thus still gaining fast convergence rate but at the price of less variable selection.

Remark 2.12. Although we assume p fixed, in practice it may still need to be estimated if no reasonable value (or upper bound) can be given. As p determines the number of selection regressions to be conducted, it has to be determined a priori and cannot be integrated in the lasso estimation. It can still be determined though by a (separate) lasso-type algorithm. For example, one may estimate (2.4) with a large initial lag length p^*, and let p be determined as the largest lag for which variables are selected, possibly also varying the lag length over variables. For this approach the hierarchical penalties of Nicholson, Wilms, et al. (2020) provide a better option than the regular lasso, as the regular lasso tends to select occasional “spurious” high lags, which would have a significant impact on the testing procedure. Alternatively one may marginalize the VAR to a collection of univariate AR(p) processes, and select the lag length by minimizing an information criterion on the residual covariance matrix. As marginalization increases the lag length, such an approach would yield a simple to compute upper bound on p.

2.4 Asymptotic Properties

In this section we derive the asymptotic properties of our method. We first present and discuss our general high-level assumption under which the properties are derived, and then state our main results.

Assumption 2. Let δ_T and Δ_T denote sequences such $\delta_T, \Delta_T \to 0$ as $T \to \infty$. Then assume that the following conditions are satisfied:

(a) **Population Eigenvalues:** Let $\boldsymbol{e}_t = (e_{1,t}, \dots, e_{N_X,t})'$, $\boldsymbol{E} = (\boldsymbol{e}_1, \dots, \boldsymbol{e}_T)'$ and $\boldsymbol{E}^{\otimes} = \boldsymbol{I}_{N_I} \otimes \boldsymbol{E}$. Define

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{GC,GC} & \boldsymbol{\Sigma}_{GC,-GC} \\ \boldsymbol{\Sigma}_{-GC,GC} & \boldsymbol{\Sigma}_{-GC,-GC} \end{bmatrix} = \begin{bmatrix} \mathbb{E}\left(\boldsymbol{x}_{GC,t}\boldsymbol{x}'_{GC,t}\right) & \mathbb{E}\left(\boldsymbol{x}_{GC,t}\boldsymbol{x}'_{-GC,t}\right) \\ \mathbb{E}\left(\boldsymbol{x}_{-GC,t}\boldsymbol{x}'_{GC,t}\right) & \mathbb{E}\left(\boldsymbol{x}_{-GC,t}\boldsymbol{x}'_{-GC,t}\right) \end{bmatrix}$$

Then there exists a constant $c_L > 0$ not depending on T and k such that $\lambda_{\min}(\boldsymbol{\Sigma}) > c_L$, where $\lambda_{\min}(\boldsymbol{\Sigma})$ denotes the minimum eigenvalue of $\boldsymbol{\Sigma}$.

(b) **Limit Behavior:** Let

$$\boldsymbol{E}^{\otimes\prime}\boldsymbol{u}_I/\sqrt{T} = \text{vec}(\boldsymbol{E}'\boldsymbol{U}_I)/\sqrt{T} = \frac{1}{\sqrt{T}}\sum_{t=p+1}^{T} \text{vec}(\boldsymbol{e}_t\boldsymbol{u}'_{I,t}) \xrightarrow{d} N(0, \boldsymbol{\Omega}),$$

$$\begin{aligned} \boldsymbol{E}'\boldsymbol{E}/T &= \frac{1}{T}\sum_{t=p+1}^{T} \boldsymbol{e}_t\boldsymbol{e}'_t \xrightarrow{p} \boldsymbol{\Sigma}_{GC|-GC} = \\ &= \boldsymbol{\Sigma}_{GC,GC} - \boldsymbol{\Sigma}_{GC,-GC}\boldsymbol{\Sigma}^{-1}_{-GC,-GC}\boldsymbol{\Sigma}_{-GC,GC}, \\ \boldsymbol{U}'_I\boldsymbol{U}_I/T &\xrightarrow{p} \boldsymbol{\Sigma}_{u,I}, \end{aligned}$$

where $\boldsymbol{\Omega} = \text{plim}_{T\to\infty}\left(\boldsymbol{E}^{\otimes\prime}\boldsymbol{u}_I\boldsymbol{u}'_I\boldsymbol{E}^{\otimes}\right)/T$.

(c) **Empirical Process:** We have with probability at least $1-\Delta_T$ that $\left\|\boldsymbol{X}'_{-GC}\boldsymbol{u}_i/\sqrt{T}\right\|_\infty \leq \gamma_T$ for all $i \in I$ and $\left\|\boldsymbol{X}'_{-GC}\boldsymbol{e}_j/\sqrt{T}\right\|_\infty \leq \gamma_T$ for all $j = 1, \ldots, N_X$, with $\boldsymbol{e}_j$ the j-th column of $\boldsymbol{E}$, for some deterministic sequence γ_T subject to the restrictions in (h).

(d) **Boundedness:** The (Granger causality) parameters of interest are bounded, that is, there exists a fixed constant $C > 0$ such that $\|\boldsymbol{\beta}_{GC}\|_1 \leq C$.

(e) **Consistency:** The initial estimators $\hat{\boldsymbol{\gamma}}_j$ are consistent in the prediction sense; specifically, with probability at least $1 - \Delta_T$ we have that

$$\left\|\boldsymbol{X}^{\otimes}_{-GC}(\hat{\boldsymbol{\gamma}}_0 - \boldsymbol{\gamma}_0)\right\|_2/\sqrt{T} \leq \delta_T T^{-1/4},$$
$$\max_{j=1,\ldots,N_X} \|\boldsymbol{X}_{-GC}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j)\|_2/\sqrt{T} \leq \delta_T T^{-1/4}.$$

(f) **Sparsity:** Let $S_j = \{m : \gamma_{m,j} \neq 0\}$ denote the sets of active variables in (2.6) and (2.7) and let $s = |S_0| + \sum_{j=1}^{N_X} |S_j|$. Let $\hat{s}$ be as defined in Algorithm 1. Then both the DGP and the initial estimators are sufficiently sparse; in particular, we have that with probability at least $1 - \Delta_T$, $\max(s, \hat{s}) \leq \bar{s}_T$ for a deterministic sequence $\bar{s}_T$ subject to the restrictions in (h).

(g) **Sparse Eigenvalues:** for any $\boldsymbol{\gamma} \in \mathbb{R}^{(K-N_J)p}$ with $\|\boldsymbol{\gamma}\|_0 \leq \bar{s}_T$, we have with probability at least $1 - \Delta_T$ that $\|\boldsymbol{\gamma}\|_2^2 \leq \left\|\boldsymbol{X}_{-GC}\boldsymbol{\gamma}/\sqrt{T}\right\|_2^2/\phi^2_{T,\min}$, where $\phi_{T,\min} > 0$ is subject to the restrictions in (h).

(h) **Rate Conditions:** The deterministic sequences $\bar{s}_T, \gamma_T$ and $\phi_{T,\min}$ introduced above satisfy the restriction $\bar{s}_T\gamma_T/\phi_{T,\min} \leq \delta_T T^{1/4}$.

Assumption 2 is a high-level assumption that allows for much flexibility on the underlying DGP and the used estimators in the first step. We now discuss each part in turn. Part (a) assumes that the minimum eigenvalue of $\boldsymbol{\Sigma}$ is bounded. This is required for application of lasso methods, as well as for the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ and the projection coefficients in (2.6) and (2.7) to exist. Part (b) assumes that a central limit theorem and weak law of large numbers hold. Essentially this require that the process is sufficiently well-behaved in terms of moments and dependence allowed. Although for convenience we assume martingale difference errors in Assumption 1, (b) holds under much weaker conditions such as mixing errors; see e.g. Davidson (1994, Chapter 14).

Part (c) is closely related to (b), but additionally controls the tail behavior of the empirical process. Results of this kind are standard in the lasso literature and can be derived using a variety of tail bounds depending on the properties of the random variables of interest, see e.g. Kock and Callot (2015) and Medeiros and Mendes (2016a) for results relevant to VAR and time series models. Of particular interest for financial applications, Masini et al. (2019, Proposition 2) show that this condition is satisfied for VAR models with general weakly dependent errors that include many popular multivariate volatility models. The boundedness assumption in (d) is not very restrictive, and with N_{GC} fixed follows directly if the parameter space of $\boldsymbol{\beta}$ is a compact set.

Part (e) imposes an appropriate consistency rate on the predictions coming from the first-stage estimator. Such prediction consistency is a standard result for lasso estimators; in particular, Wong et al. (2020) obtain it for a very general class of VAR models allowing for conditional heteroskedasticity and dependence in the error terms. Adamek et al. (2020) derive consistency of the lasso under misspecified time series models, and show that their setting covers (among others) the first-step regressions of the relevant predictors in $\boldsymbol{X}_{GC}$ on the other regressions, which are inherently misspecified in a VAR setup due to the missing lags; see their Remark 3 for further details.

Next to consistency, we also require sparsity of the DGP and the estimator, as controlled by part (f). The assumption of exact sparsity in the DGP for the initial regressions can be relaxed to approximate sparsity as in Belloni, Chernozhukov, and Hansen (2014b). For the sake of expositional clarity we do not work under that assumption here but stick to the simpler exact sparsity. Sparsity of the first-stage estimator is needed in our framework as we perform OLS on the selected variables from the first-stage regressions. If the selected variables are not sparse enough, too many variables will be selected for OLS to be feasible. Sparsity of lasso estimators is analysed in Belloni and Chernozhukov (2013), while Kock and Callot (2015) and Medeiros and Mendes (2016a) provide results for adaptive lasso for time series. Importantly, we do not require consistent model selection; the selection method used is allowed to make "persistent" mistakes, allowing for both variables to be incorrectly included and relevant variables to be missed, as long as the estimator remains sufficiently sparse and consistency is guaranteed. Unlike Belloni, Chernozhukov, and Hansen (2014b), we allow for the order of sparsity of the estimator to differ from the true sparsity thereby opening the way for conservative selection procedures.

Given the assumptions above, the eigenvalue assumption in (g) becomes almost superfluous, as it is generally needed to establish (e) and (f) for lasso-type estimators; see e.g. Belloni and Chernozhukov (2013) and Medeiros and Mendes (2016a) for details. It requires that for sufficiently sparse vectors, the eigenvalues of the subset of the Gram matrix corresponding to their non-zero support do not decrease to zero too fast. Such assumptions are standard in the lasso literature in various guises as *restricted eigenvalue* conditions, and can typically be derived by making similar conditions on the population covariance matrix $\boldsymbol{\Sigma}_{-GC,-GC}$ coupled with a convergence result of the Gram matrix $\boldsymbol{X}'_{-GC}\boldsymbol{X}_{GC}$ to $\boldsymbol{\Sigma}_{-GC,-GC}$. Basu, Shojaie, et al. (2015), Masini et al. (2019) and Wong et al. (2020) establish the plausibility of such restricted eigenvalue conditions for various VAR models. We state the condition here explicitly as it is needed directly in the proofs.

Finally, note that the restrictions on tail behavior (via γ_T), sparsity (via $\bar{s}_T$) and minimum eigenvalues (via $\phi_{T,\min}$) are meaningless if no rates on these sequences are imposed. Part (h) therefore is the key part which connects all assumptions with explicit rates needed for the validity of the PDS method. The restrictions here represents a trade-off between sparsity, thickness of tails and minimum eigenvalues. For example, if, as often assumed $\phi_{T,\min}$ is fixed and $\boldsymbol{u}_t$ is Gaussian, tails are sufficiently thin that γ_T can be chosen as roughly the order of $\sqrt{\ln(K^2 p)}$ (cf. Kock and Callot, 2015, Lemma 4), leaving room for either almost exponentially large K relative to T, or a fairly non-sparse model. On the other hand, if only m moments of $\boldsymbol{u}_t$ exist, γ_t should be taken roughly of the order $(K^2 p)^{2/m}$ (Masini et al., 2019, Lemma 2), requiring polynomial growth of K compared to T and sparser models.

The most restrictive and crucial assumption needed on the underlying DGP for satisfying Assumption 2 is the sparsity of the underlying DGP formulated in part (f). The plausibility of this assumption highly depends on the specific application. In many financial applications sparsity (or its approximate version) is natural, for example in portfolio selection when the number of assets is large and the estimation of high-dimensional volatility matrices in financial risk assessment (see Fan, Lv, et al. (2011) for an overview), as well as in our investigation of Granger causality in networks of realized volatilities in Section 2.6. The volatility of one particular stock is likely to have specific channels of contagion rather than affecting the whole stock market at the same time. Shocks to one asset therefore likely propagate through the system via specific channels, which corresponds to sparse lag polynomials. One might worry about systemic shocks affecting many assets; however, the dense covariance matrix $\boldsymbol{\Sigma}_u$ can accommodate simultaneous common shocks. Moreover, the dynamic of such shocks can generally well be captured through a sparse combination of the most important and most affected assets. Similarly, in macroeconomic applications it has been found that a few important variables can capture the effects of unobserved common factors, leading sparse models to perform as well as common factors (De Mol et al., 2008; Smeekes and Wijler, 2018).

We are now ready to state our main asymptotic result of this section in Theorem 2.1 which establishes the asymptotic normality of the post-lasso (generalized) least squares estimator. Here we slightly deviate from the LM test in Algorithm 1; after the double selection procedure carried out in Step [1], we regress the transformed outcome variables $\widetilde{\boldsymbol{y}}_t = (\boldsymbol{G}_T \otimes \boldsymbol{I}_T)\boldsymbol{y}_I$ on both the Granger causing $\widetilde{\boldsymbol{X}}^{\otimes}_{GC} = (\boldsymbol{G}_T \otimes \boldsymbol{I}_T)\boldsymbol{X}^{\otimes}_{GC}$ and selected variables $\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}} = (\boldsymbol{G}_T \otimes \boldsymbol{I}_T)\boldsymbol{X}^{\otimes}_{\hat{S}^{\otimes}}$

$$\widetilde{\boldsymbol{y}}_I = \widetilde{\boldsymbol{X}}^{\otimes}_{GC}\boldsymbol{\beta}^{\text{PDS}}_{GC} + \widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}}\boldsymbol{\beta}^{\text{PDS}}_{\hat{S}^{\otimes}} + \widetilde{\boldsymbol{u}}_I. \tag{2.8}$$

The transformation by the matrix $\boldsymbol{G}_T$ allows for the GLS esitmation needed in the LM procedure by taking $\boldsymbol{G}_T = \hat{\boldsymbol{\Sigma}}^{-1/2}_{u,I}$, while OLS is performed with $\boldsymbol{G}_T = \boldsymbol{I}_{N_I}$. In the latter case the theorem provides the foundation for the Wald test discussed in Remark 2.11 (minus the required variance estimation for that test). We state this result separately as it is interesting in its own right, and can be used to establish validity of other tests such as the Wald test.

Theorem 2.1. *Let $\hat{\boldsymbol{\beta}}^{\text{PDS}}_{GC}$ denote the OLS estimator of $\boldsymbol{\beta}^{\text{PDS}}_{GC}$ in 2.8. Let $\boldsymbol{G}_T$ be any matrix satisfying with probability at least $1-\Delta_T$ that $0 < c_1 \leq \lambda_{\min}(\boldsymbol{G}'_T\boldsymbol{G}_T) \leq \|\boldsymbol{G}'_T\boldsymbol{G}_T\|_{\max} \leq c_2 < \infty$, where c_1, c_2 are constants not depending on T. Then, uniformly in all DGPs that satisfy Assumption 2, we have as $T \to \infty$,*

$$\sqrt{T}(\hat{\boldsymbol{\beta}}^{\text{PDS}}_{GC} - \boldsymbol{\beta}_{GC}) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, (\boldsymbol{G}'\boldsymbol{G} \otimes \boldsymbol{\Sigma}_{GC|-GC})^{-1}\boldsymbol{\Omega}_{\boldsymbol{G}}(\boldsymbol{G}'\boldsymbol{G} \otimes \boldsymbol{\Sigma}_{GC|-GC})^{-1}\right),$$

where $\boldsymbol{\Omega}_{\boldsymbol{G}} = \operatorname{plim}_{T\to\infty}\left[(\boldsymbol{G}'_T\boldsymbol{G}_T \otimes \boldsymbol{E}')\boldsymbol{u}_I\boldsymbol{u}'_I(\boldsymbol{G}'_T\boldsymbol{G}_T \otimes \boldsymbol{E})\right]/T.$

Theorem 2.1 establishes the asymptotic normality of the post-double-selection OLS estimators. The statement ‘uniformly in all DGPs that satisfy Assumption 2’ should be interpreted as the theorem holding uniformly over a parameter space that is defined such that Assumption 2 holds for all parameters in that parameter space. Importantly, no beta-min conditions on the smallest magnitude of parameters are required, thus alleviating the post-selection inference problem. We refer to Comments 3.4 and 3.5 in Belloni, Chernozhukov, and Hansen (2014b) for further details regarding the uniformity. The limit distribution of the LM test now follows straightforwardly from Theorem 2.1, and is stated in the corollary below.

Theorem 2.2. *Let $\boldsymbol{\beta}_{GC} = \mathbf{0}$. Then, uniformly in all DGPs that satisfy Assumption 2 and for which $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{u,I} \otimes \boldsymbol{\Sigma}_{GC|-GC}$, we have that*

$$LM \xrightarrow{d} \chi^2_{N_{GC}} \qquad \text{as } T \to \infty.$$

Theorem 2.2 establishes the limiting distribution of the PDS-LM test under an additional condition on the (co)variances of the partial regression errors, which is satisfied if the errors are iid. To allow for heteroskedaticity the LM test has to be modified, which would only lead to more cumbersome proofs without adding any novelty specific to the high-dimensional case. Therefore we focus on the homoskedastic case here, although we do consider a heteroskedasticity-robust version of the test in the volatility application in Section 2.6.[5]

2.5 Monte-Carlo Simulations

We now evaluate the finite-sample performance of our proposed Granger causality test. We consider three Data Generating Processes (DGPs) inspired by Kock and Callot (2015):

$$\text{DGP1:} \quad \boldsymbol{y}_t = \begin{bmatrix} 0.5 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0.5 \end{bmatrix} \boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t,$$

$$\text{DGP2:} \quad \boldsymbol{y}_t = \begin{bmatrix} (-1)^{|i-j|} a^{|i-j|+1} & \dots & (-1)^{|i-j|} a^{|i-j|+1} \\ \vdots & \ddots & \vdots \\ (-1)^{|i-j|} a^{|i-j|+1} & \dots & (-1)^{|i-j|} a^{|i-j|+1} \end{bmatrix} \boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t,$$

$$\text{DGP3:} \quad \boldsymbol{y}_t = \begin{bmatrix} \boldsymbol{A} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \boldsymbol{A} \end{bmatrix} \boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t \quad \text{with } \underbrace{\boldsymbol{A}}_{5\times 5} = \begin{bmatrix} 0.15 & \cdots & 0.15 \\ \vdots & \ddots & \vdots \\ 0.15 & \cdots & 0.15 \end{bmatrix},$$

[5] Note that this is no different for the Wald test, for which the variance estimation has to be adjusted as well.

where $a = 0.4$ in DGP2. The diagonal VAR in DGP1 respects the sparsity assumption while in DGP2 the entries are set to decrease exponentially fast in the distance from the main diagonal and hence the sparsity assumption is not met. DGP2 could be empirically motivated by looking e.g. at financial interconnectedness. Financial institution, such as banks, lend to and borrow from one another becoming interconnected through interbank credit exposures. The financial distress experienced by one bank is likely to be most heavily transmitted the closer the connections are as well as less transmitted, the weaker the connections. DGP3 is a block-diagonal system. Such a structure is motivated by e.g., typical quarterly macroeconomic models capturing business cycle dynamic and monetary and fiscal policy effects. One such example is DSGE models, where the dynamic of the economy through time is monitored on quarterly frequency. Note that as written above, DGP1 satisfies the null of no Granger causality from unit 2 to 1, while DGP2 and DGP3 do not. Therefore, we adapt DGP 1 for the power analysis by setting the coefficient in position $(2, 1)$ equal to 0.2. Conversely, we set the same coefficient equal to zero for DGP2 and DGP3 for the size analysis.

We choose our series of interest as $I = \{2\}$ and $J = \{1\}$, thereby focusing on the case where we have single variables of interest for both elements of the test. Here we consider for simplicity $p = 1$ lag, namely the same lag-length as in the DGPs, so $j = 1$. The equation of interest can then be written as

$$y_{2,t} = \beta_{GC} y_{1,t-1} + \sum_{j=2}^{K} \beta_j y_{j,t-1} + \epsilon_{2,t}.$$

Hence, for each DGP we test $H_0 : \beta_{GC} = 0$ against $H_1 : \beta_{GC} \neq 0$ using our proposed PDS-LM test.

Table 2.1 reports the size and power of the test for 1000 replications by using different combinations of time series length $T = (50, 100, 200, 500)$ and number of variables in the system $K = (10, 20, 50, 100)$ and a fixed lag-length $p = 1$. All the rejection frequencies are reported using a burn-in period of fifty observations. For each scenario, AIC, BIC and EBIC are compared with the theoretical choice of the tuning parameter λ^{th} and time series cross validation λ^{TSCV} as described in Subsection 2.3.3.

Simulations are also reported for different types of covariance matrices of the error terms. We employ a Toepliz-version for calculating the covariance matrix as $\Sigma_{i,j} = \rho^{|i-j|}$ by using two scenarios of correlation: $\rho = (0, 0.7)$. The first case corresponds to no correlation, and is equivalent to set $\boldsymbol{\Sigma} = \boldsymbol{I}_K$.

In the Appendix we provide some additional simulation results. First, Table 2.2 reports the simulation results for all three DGPs using $\Sigma_{i,j} = 0.7^{|i-j|}$. Second, we investigate the Wald version of our test in Table 2.3. Third, in Table 2.4 we investigate the effects of miss-specification of the lag length by estimating the over-specified VAR($p + 1$) instead of the true-order VAR(p).[6] Fourth, in Table 2.5 we report the results for the size of a bivariate Granger causality test for a non-sparse DGP when using a standard Wald (F) test. This test is obviously sensitive to omitted variable bias, and our goal is to demonstrate its effect. Finally, although all results reported here use the finite sample correction in Step 3b of the algorithm, we also investigated the differences with Step 3a. We comment on these results in the next subsection.

Our proposed approach shows a good performance in terms of size and (unadjusted) power for all DGPs considered. Both for the setting of no correlation and high correlation of errors, sizes are in the vicinity of 5% and power is increasing with the sample size T.

Only moderate size distortion is visible in large systems for small samples (e.g. $K \geq 50,\ T = 50$). As expected, the test procedure works remarkably well for the sparse DGP1 in high dimensions. However, size properties under the non-sparse DGP2 do not deviate much from

[6] For both the Wald test and the over-specified VAR(p+1) we report the simulations for $\Sigma_{i,j} = 0.7^{|i-j|}$ and DGP1 only. Results for the other DGPs are available upon request.

Table 2.1: Simulation results for the PDS-LM Granger causality test

DGP	Size/Power	ρ	T	50					100					200					500				
			K	AIC	BIC	EBIC	λ^{th}	λ^{TSCV}	AIC	BIC	EBIC	λ^{th}	λ^{TSCV}	AIC	BIC	EBIC	λ^{th}	λ^{TSCV}	AIC	BIC	EBIC	λ^{th}	λ^{TSCV}
1	Size	0	10	6.7	6.1	5.3	6.8	6.9	6.7	6.1	5.9	6.5	6.4	5.5	4.5	4.5	4.7	5.4	4.3	4.1	4.3	4.2	4.3
			20	7.3	6.0	6.4	5.9	6.2	6.0	4.7	4.7	5.5	5.8	4.0	4.7	5.4	4.3	4.5	4.3	4.1	4.3	4.1	4.3
			50	7.0	5.9	5.7	6.4	5.5	7.4	6.4	6.4	6.2	6.6	6.9	5.9	5.9	6.2	6.1	6.8	6.5	6.7	6.8	6.7
			100	7.4	7.2	6.6	NA	6.8	7.2	4.9	4.9	5.9	5.4	6.5	4.7	4.9	5.0	5.2	5.8	3.9	4.0	5.0	4.1
1	Power	0	10	30.0	31.2	33.3	30.4	30.9	58.3	58.9	60.7	58.5	57.4	89.1	89.3	89.6	89.2	89.5	99.9	99.9	99.9	99.9	99.9
			20	22.8	26.4	30.2	25.8	24.2	52.8	55.1	57.4	53.9	54.5	85.6	88.0	89.0	86.6	86.1	99.9	100	100	99.9	99.9
			50	13.6	24.3	33.3	17.9	18.7	38.7	53.8	59.0	46.4	45.5	78.2	85.8	87.0	81.3	80.2	99.9	99.9	99.9	99.9	99.9
			100	10.4	20.3	32.1	NA	18.7	20.4	51.9	57.0	31.2	35.8	61.6	85.4	86.7	74.1	73.7	99.5	100	100	99.8	99.7
2	Size	0	10	6.2	5.5	6.1	6.1	5.9	5.0	4.9	5.4	5.0	5.0	4.7	4.6	4.9	4.6	5.2	3.9	3.9	3.9	3.7	3.7
			20	7.8	5.5	6.0	5.8	6.5	5.4	4.6	4.7	4.8	5.3	5.9	5.9	5.4	5.7	6.0	4.7	3.8	4.3	4.6	4.7
			50	7.3	5.9	4.7	6.0	7.1	7.9	6.3	6.5	7.3	6.5	7.3	6.1	6.0	6.6	6.6	6.3	6.6	6.4	6.1	6.0
			100	6.1	6.9	5.5	NA	6.6	6.8	5.1	6.0	5.3	4.8	5.4	5.5	5.8	4.6	4.9	6.1	4.3	5.0	4.9	4.5
2	Power	0	10	18.0	19.7	21.3	18.2	17.0	37.9	39.3	40.4	38.0	38.4	64.7	64.7	66.9	64.8	65.1	97.4	97.4	97.5	97.4	97.4
			20	16.0	19.6	24.6	18.8	16.9	35.4	39.8	44.4	37.4	36.7	64.4	67.4	69.7	66.0	65.0	97.2	97.5	97.5	97.5	97.4
			50	8.6	15.2	21.7	12.8	13.1	25.0	36.1	43.2	32.6	31.2	57.0	66.4	71.9	61.8	61.1	95.0	96.2	96.8	96.1	95.4
			100	9.2	14.1	25.1	NA	10.1	15.1	34.9	45.9	25.9	25.8	44.8	65.1	74.7	58.0	57.8	94.7	97.3	97.7	96.3	96.5
3	Size	0	10	5.2	5.0	5.6	5.7	4.6	5.6	4.9	5.7	5.7	6.2	4.0	4.1	6.1	4.1	4.4	4.1	4.0	3.8	3.9	4.1
			20	4.2	5.1	5.7	5.6	4.9	4.3	4.3	7.2	4.2	3.9	5.2	5.6	9.4	4.7	4.6	4.7	4.4	4.5	4.5	4.8
			50	7.5	6.3	7.4	6.9	6.6	6.4	7.0	9.5	5.9	5.5	6.9	6.9	12.4	6.0	6.6	4.9	5.3	6.6	5.3	5.4
			100	7.1	6.7	8.3	NA	7.0	6.2	5.7	8.3	5.7	6.0	4.7	6.2	10.7	4.2	4.7	4.4	5.1	6.5	5.1	4.7
3	Power	0	10	15.4	20.0	23.4	16.2	16.1	31.5	36.4	44.3	32.6	31.4	58.4	61.3	63.7	58.9	59.3	95.2	95.6	95.7	95.5	95.3
			20	13.4	18.7	26.0	13.8	13.8	29.5	37.0	47.8	30.0	29.8	56.5	62.0	69.8	56.8	55.5	94.0	94.6	94.8	94.4	94.3
			50	11.4	20.7	28.0	12.9	10.6	24.1	39.8	52.3	26.9	27.6	50.3	59.5	73.6	52.1	51.8	91.1	92.7	93.4	92.3	90.9
			100	7.9	18.7	26.9	NA	14.1	13.7	42.6	55.0	20.2	22.2	41.4	62.0	75.2	44.9	44.8	90.0	94.0	94.8	91.4	90.2

Notes: Size and Power for the different DGPs described in Section 2.5 are reported for 1000 replications. $T = (50, 100, 200, 500)$ is the time series length, $K = (10, 20, 50, 100)$ the number of variables in the system, the lag-length is fixed to $p = 1$. ρ indicates the correlation employed to simulate the time series with the Toeplitz covariance matrix. The different choices of the tuning parameter λ are reported as: AIC, BIC, EBIC for information criteria, λ^{th} for the theoretical plug-in and TSCV for time series cross-validation as explained in Section 2.5.

its sparse counterpart, although for both DGP2 and DGP3 we do observe a slight deterioration of size when the dimension of the system increases.

Interestingly, the three different information criteria show substantially different behavior. EBIC, due to its very stringent nature, tends to perform well only in very large systems, while it is essentially equal to a bivariate Granger causality test in small systems. We have to add though that the good performance of AIC in particular is somewhat inflated by the imposed lower bound on the penalty; unreported simulations show that without the lower bound AIC performs significantly worse, often selecting too many variables rendering the post-OLS estimation infeasible. The one advantage of using EBIC as information criterion to tune λ in the $K >> T$ settings when T is small (e.g. $T = 50, 100$) is the possibility to avoid the lower bound on the penalty. However, since this comes at a price of more size distortion, we recommend the use of BIC instead, along with the lower bound on the penalty. When comparing the different choices of the tuning parameter we can narrow down the best performing ones (in terms of size and power) to BIC and λ^{th}. However, in terms of computational time, estimating the tuning parameter using information criteria is considerably faster.

Comparing our test to the bivariate VAR in Table 2.5, it is clear that our proposed PDS-LM is very robust to omitted variable bias, unlike the bivariate test, whose size distortions increase with both the sample size and the number of variables, with sizes of 45% observed for the sample sizes we consider in our application in Section 2.6. There we will also further elaborate on this difference between our method and the bivariate test. Table 2.1 shows that for sample sizes smaller than $T = 500$, rarely the power exceeds 90%. However, one must keep in mind that the powers are not size-adjusted, and thus the high reported power of the low-dimensional test is an artefact of the huge size distortions rather than genuine power. It also seems unreasonable to expect that PDS-LM test has vey high power if T is small; we are still considering large systems with many parameters to estimate, and there seems to be no way around this if one desires to test Granger causality in large systems with many (control) variables. In that sense we may fully expect the bivariate test to also have higher size-corrected power; yet with all its disadvantages and sensitivity to omitted variables this is not a good comparison. All in all, we believe our test still has sufficiently adequate power properties to be useful in practice.

The results of robustness to misspecification of the lag length order with $p = 2$ instead of $p = 1$ are reported in Table 2.3 in Appendix B. As the size distortions across the range of considered DGPs are only marginally higher for large K and T comparatively small, the test appears to be quite robust to this misspecification. Again, BIC seems to be the best choice for tuning the penalty for all DGPs. Unreported simulations (available upon request) further show that the finite sample adjustment for the test performed in Step 3b of the algorithm is able to substantially reduce size distortions in small samples compared to the asymptotic version of Step 3a.

2.6 Networks in Realized Volatilities

2.6.1 Realized Variances

We first investigate the volatility transmission in stock return prices using the daily realized variances of 30 US assets.[7] Both the computational simplicity and the theoretical foundations make realized volatility measures (realized variance, bi-power variation, median realized variance, etc.) very attractive among practitioners and academics for modelling time varying volatilities and monitoring financial risk. We have considered 10-minute realized variances

$$RV10_t \equiv \sum_{j=1}^{M} r_{j,t}^2, \qquad r_{j,t} = \ln P_{j,t} - \ln P_{j-1,t}, \tag{2.9}$$

[7] We would like to thank Marcelo C. Medeiros for providing us with the high frequency data on stock prices that we have used to construct the realized variances. See Table 2.6 for the stocks considered. The R package HDGCvar is available on the GitHub page of the corresponding author (`https://github.com/Marga8`).

using $j = 1, \dots, M$ intraday 10 minutes stock prices $P_{j,t}$. We consider 10 minute returns as this is the frequency that minimizes for our sample the microstructure noise (McAleer and Medeiros (2008)).[8] We investigate the period from March 2008 until February 2017 (2236 trading days).

Given the time series of realized volatilities as defined in (2.9), we employ a multivariate version of the heterogeneous autoregressive model (VHAR) of Corsi (2009) to model their joint behavior (see also Cubadda et al. (2019)). To formally define the VHAR model, we log-transform the series and we stack the logarithmic RV into a vector y_t. The VHAR specification is given by the following model:

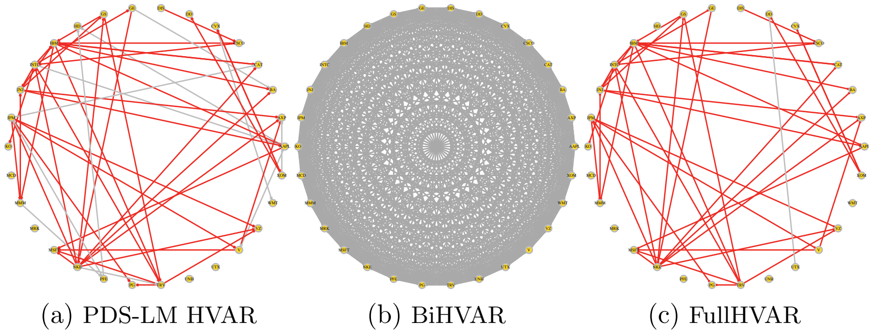
$$\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{B}^{(1)}\boldsymbol{y}_{t-1} + \boldsymbol{B}^{(2)}\boldsymbol{y}_{t-1}^{(week)} + \boldsymbol{B}^{(3)}\boldsymbol{y}_{t-1}^{(month)} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{y}_t^{(week)} = \frac{1}{5}\sum_{j=0}^{4}\boldsymbol{y}_{t-j}$ and $\boldsymbol{y}_t^{(month)} = \frac{1}{22}\sum_{j=0}^{21}\boldsymbol{y}_{t-j}$ are the vectors containing the average volatility over the last 5 (week) and 22 (month) days. Granger causality in this context represents contagion, or spillover, of volatility from one asset to another. To test for the null hypothesis of no Granger causality / no volatility spillovers from $y_{k,t}$ to $y_{i,t}$ against the alternative of spillovers, we test

$$H_0 : \beta_{i,k}^{(1)} = \beta_{i,k}^{(2)} = \beta_{i,k}^{(3)} = 0 \qquad \text{vs.} \qquad H_1 : \beta_{i,k}^{(1)}, \beta_{i,k}^{(2)}, \beta_{i,k}^{(3)} \neq 0,$$

where $\beta_{i,k}^{(1)}$ is the (i,k)-th element of $B^{(1)}$. We perform this test for every (i,k)-pair to obtain the full 29×29 network of spillover effects. As heteroskedasticity is likely present in these data, we robustify the PDS-LM procedure by implementing the heteroskedasticity-robust LM test such as for example described in Wooldridge (2015, Ch. 8). The

[8] To determine the optimal frequency, we computed realized variances using different frequencies of 1, 5, 10, 15, 30, 65 and 130 minutes, in addition to the estimation using daily returns. The latter estimation has the advantage of being unbiased but the drawback of being very noisy (Pooter et al. (2008)). To find an optimal trade-off between bias and variance (Martens, 2004, see e.g)), mean, variances and mean squared errors (MSE) were computed for each estimation frequency in a similar way as Pooter et al. (2008), and it was found that the frequency of 10 minutes minimizes the MSE.



(a) PDS-LM HVAR (b) BiHVAR (c) FullHVAR

Figure 2.1: Spillover networks for the full sample period

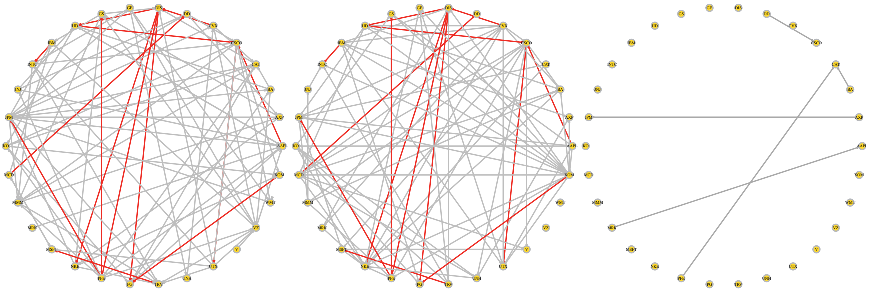
full algorithm for the heteroskedasticity-robust PDS-LM test is given in Appendix C.[9]

We now report the results of our spillover tests for the volatility network. We use BIC to select the tuning parameter of the lasso, and perform the Granger causality tests with a 1% significance level.[10] Figure 2.1 reports the transmission networks of volatilities estimated with the high-dimensional HVAR (PDS-LM HVAR), bivariate Granger causality tests (BiHVAR) for each pair of stocks, Granger causality tests from a full-system VAR (FullHVAR). The latter is feasible because of our large time series dimension with $T = 2236$. For all methods we consider heteroskedasticity-robust variants.

While our PDS-LM method identifies a volatility transmission network which consists of 54 connections and the FullHVAR test picks up 44

[9] In the presence of heteroskedasticity, one might prefer the Wald version of the test, as this can be corrected in the standard way by using heteroskedastiсty-robust standard errors. Empirically we found hardly any differences between the LM and Wald versions.

[10] We do not perform a correction for multiple testing, as this would only qualitatively affect our results. Moreover, our goal is not to identify exactly the set of spillovers, but to get a feeling of the relations between two variables at a time. As such, we believe a multiple testing correction is not needed, though it can be easily implemented.

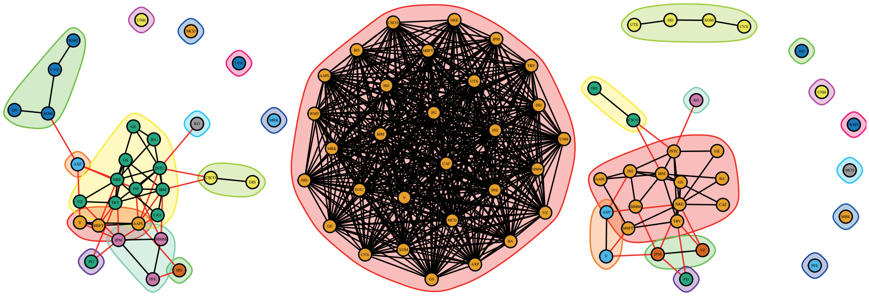


(a) PDS-LM HVAR (b) BiHVAR (c) FullHVAR

Figure 2.2: Spillover networks for the 2016-2017 sample period

connections, the BiHVAR tests detect a network consisting of 803 connections. These differences between our PDS-LM HVAR and the BiHVAR results are in line with our simulation results, confirming that bivariate Granger causality testing in VAR models is seriously affected by omitted variable bias in high-dimensional systems. Given the huge sample size ($T = 2236$) relative to the number of stocks, the FullHVAR is a feasible option, and it is reassuring how similar our PDS-LM HVAR performs compared to the FullHVAR. The similarity is visualized in Figures 2.1(a) and 2.1(c), where the connections picked by both methods are highlighted in red. Of the 54 spillovers identified by the PDS-LM HVAR, 43 are also identified by the FullHVAR, while only 1 of the identified spillovers by the FullHVAR is not picked up by the PDS-LM HVAR.

We also consider a shorter time span, namely the period 2016-2017. Considering a shorter time period makes it more likely that relations remain stable over time. In particular, the chosen period avoids two major events that occurred previously and caused substantial instability on financial markets, namely the global financial crisis of 2008 and the U.S. debt-ceiling crisis of 2011 (Baker et al., 2019). It also allows us to study the performance differences among the three methods in a smaller sample of $T = 284$ trading days, where the FullHVAR suffers from the



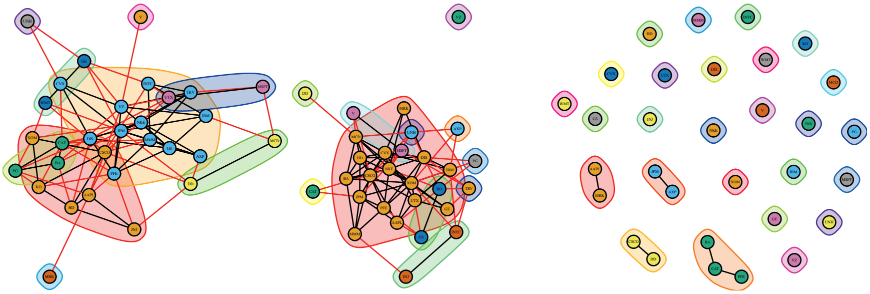
(a) PDS-LM HVAR (b) BiHVAR (c) FullHVAR

Figure 2.3: Volatility clusters for the full sample period

curse of dimensionality. We present the results for the PDS-LM HVAR, BiHVAR and FullHVAR in Figure 2.2. The number of significant transmissions is 91 for the PDS-LM HVAR, 85 for the BiHVAR and only 5 for the FullHVAR. Hence, the FullHVAR breaks down in this setting due to the small sample size and curse of dimensionality. On the other hand, while superficially the PDS-LM HVAR and the BiHVAR appear to perform similarly, they identify mostly different spillovers. The red lines in Figures 2.2(a) and 2.2(b) show the common connections, which are only 14 out of 91 for the PDS-LM HVAR (85 for the BiHVAR).

As a next step, we use our identified networks to find clusters of closely connected stocks, or communities as they are called in graph theory. Communities are groups of densely connected nodes with fewer connections across groups. In order to represent volatility spillover communities in the graph we use the Newman and Girvan (2004) algorithm based on edge-betweenness. The edge betweenness for edge e is defined as $\sum_{s,t\neq e} \frac{\sigma_{st}(e)}{\sigma_{st}}$, where σ_{st} is total number of shortest paths from node s to node t and $\sigma_{st}(e)$ is the number of shortest paths passing through e. The edge with the highest betweenness is sequentially removed and the betweenness is recalculated at each step until the best partitioning of the network is found.

Figure 2.3 reports the graphs of the clustered network for the full sample



(a) PDS-LM HVAR (b) BiHVAR (c) FullHVAR

Figure 2.4: Volatility clusters for the 2016-2017 sample period

analysis for the PDS-LM VAR, BiHVAR and FullHVAR respectively. The results for the PDS-LM VAR and FullHVAR show similar spillover clustering behavior, as expected. One large big-industry cluster, containing – among others – assets such as Johnson & Johnson (J&J), IBM, Nike (NKE) and Intel (INTC) dominates the picture being surrounded by small clusters containing 1 to 4 stocks. The PDS-LM VAR and FullHVAR resepctively identify 4 and 6 isolated stocks, which do not have any connections to others. Instead, the BiHVAR finds one single cluster containing all stocks. This reinforces our finding that bivariate Granger causality testing is not informative in high-dimensional systems.

The clusters for the analysis done on the smaller 2016-2017 sample are reported in Figures 2.4a, 2.4b and 2.4c. The patterns highlighted in the spillover network graphs re-occur in the clusters. PDS-LM HVAR in Figure 2.4a picks up two main clusters of volatility spillovers containing 12 and 6 assets. In addition, four medium size clusters and three single-stock clusters are found. The difference between PDS-LM HVAR and BiHVAR is also reflected in the identified clusters. BiHVAR in Figure 2.4b shows only one big conglomerate cluster of stocks linked to three two-stock clusters and 6 single-stock clusters. Finally, the breakdown of FullHVAR shows clearly in the non-informative, mostly unconnected single-stock clusters in Figure 2.4c.

2.6.2 Realized Variances & Covariances

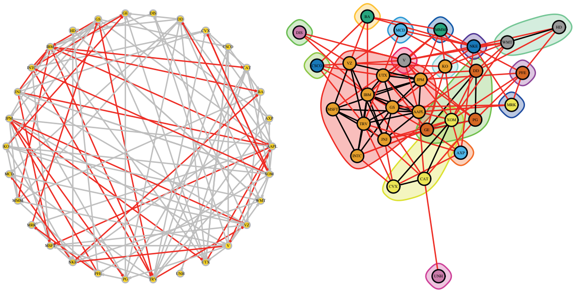
In this subsection we extend our investigation to allow for spillovers from realized correlations to variances. While our application in Section 2.6.1 was only high-dimensional when we considered the shorter subsample, including correlations, which are of the order K^2, put a significantly larger strain on estimation, making the standard full VAR no option. As elaborated later, it appears quite reasonable to expect changing correlations to also have an affect on the volatilities. By ignoring these in Section 2.6.1, we are exposing ourselves again to a potential omitted variable bias. However, our method can directly incorporate these, as we demonstrate here.

While we remain mostly interested in contagion between the 30 realized volatilities, we add the $\frac{30\times29}{2} = 435$ realized correlations between all these assets as control variables. By maintaining our focus on the relations between the variances, our results are directly comparable to Section 2.6.1 and can be interpreted by assessing how the network changes when the correlations are added as controls in the VAR. Moreover, it also avoids the difficulties of trying to visualize the results from all $(30 \times 435)^2$ possible connections in the large VAR considered here. In the same way that the realized variances employ high frequency data to estimate the integrated variance, the realized covariance (RC hereafter) estimates the integrated covariance of a multivariate diffusion process. Working with the full RC time-varying matrix is important for portfolio allocation and risk management. For a set of n intra-day asset returns at day t observed at $j = 1, \ldots, M$ stacked in a column vector $\boldsymbol{r}_{j,t}$, the realized covariance is obtained such as $\boldsymbol{RC}_t = \sum_{j=1}^{M} \boldsymbol{r}_{j,t}\boldsymbol{r}'_{j,t}$. Note that the realized variances are obviously on the diagonal of RC and that the RC matrix is positive definite when $M > n$, namely when the number of high frequency observations per day is larger than the number of series. We investigate the same period as before and construct 10-minute realized covariances. Several studies have also proposed a Lasso framework on RC, see for instance Callot et al. (2017) and Brito et al. (2018), although their focus is more on portfolio allocation and forecasting.

There are two main ways to work with the RC matrix. The first approach stacks realized variances and covariances in a single vector. For instance, Bauer and Vorkink (2011) consider the matrix log transformation of RC_t series, a matrix that they call the log-space volatility. The drawback of that log transform is that the interpretation of the original series, in our case the volatilities, is lost as the combinations involve nonlinear transforms of both realized variances and covariances. This is not compatible with the aim of this chapter.

The second approach uses the log realized volatilities and the correlations separately, as done by for instance Oh and Patton (2016). The underlying idea, following the DCC model of Engle (2002), is to decompose $RC_t^{(d)} = D_t^{(d)} R_t^{(d)} D_t^{(d)}$ with $D_t^{(d)}$ a diagonal matrix with the square root of the individual realized variance and $R_t^{(d)}$ the realized correlation matrix. Oh and Patton (2016) use the HVAR model structure for each realized volatilities, they consequently assume no Granger causality across volatilities.

We propose something which is, to some extent, in between these two approaches. We look at two separate objects as in the DCC model, but stack the log of the realized variances $\boldsymbol{y}_{1t}^{(d)\prime}$ and z-transforms $\boldsymbol{y}_{2t}^{(d)} = \operatorname{arc\,tanh}\left(\operatorname{vech}(\boldsymbol{R}_t^{(d)})\right)$ of the realized correlations in a larger vector $\boldsymbol{y}_t^{(d)} = (\boldsymbol{y}_{1t}^{(d)\prime}, \boldsymbol{y}_{2t}^{(d)\prime})'$, where $\boldsymbol{y}_{1t}^{(d)}$ is 1×30 and $\boldsymbol{y}_{2t}^{(d)}$ is 1×435, on which we estimate a HVAR of dimension 465. In this HVAR each of the 465 equations depends on 1395 dynamic parameters plus the constant. We focus on the 30 equations corresponding to $\boldsymbol{y}_{1t}^{(d)}$ volatilities and consequently the bivariate causalities between these realized volatilities as in the previous section. Figure 2.5a reports a total of 113 connections, which is about twice the connections in Figure 2.1a.



(a) PDS-LM HVAR (b) PDS-LM HVAR

Figure 2.5: Spillover network and volatility clusters

In red we highlighted the 31 common connections with Figure 2.1a. Interestingly, adding more variables therefore allows us to uncover more relations. It seems that this allows us to uncover partial effects that were previously obscured by counteracting effects of the correlations. Importantly, the number of connections is still far less than compared to the BiHVAR in Figure 2.1b, and the PDS-LM HVAR is still able to deliver a clear picture of the causal connections when the system considered is high-dimensional. While the different connections found here obviously also lead to a different clustering, Figure 2.5b shows that the clustering is quite similar, certainly regarding qualitative conclusions.

2.7 Conclusion

We propose an LM test in order to test for Granger causality in high-dimensional VAR models. We employ a post-double selection procedure using the lasso to select the set of relevant covariates in the system. The double selection step allows to substantially reduce the omitted variable bias and thereby allowing for valid post-selection inference on the parameters.

We provide an extensive simulation study to evaluate the performance of our method in finite samples, paying particular attention to the tuning of the penalty parameter. We compare different information criteria, time series cross-validation and a plug-in method based on theoretical arguments, and find that generally BIC and the theoretically tuned penalty perform best. However, to use information criteria in systems with a significantly larger number of variables than observations, a lower bound on the penalty parameter is needed to prevent too many variables being selected. The simulations also show that, when properly tuned, our proposed PDS-LM test attains good results both for size and power under different DGPs. Especially, it is shown to be robust both to non-sparse settings as well as to lag-length overspecification.

We also empirically investigate the usefulness of our method in a study where we apply our PDS-LM method to a high-dimensional VHAR process in order to construct a contagion network of volatility spillovers for 30 large capital stocks, also accounting for effects from changing correlations. We find that by increasing the information set through considering a high-dimensional VAR model instead of bivariate models, we are able to obtain more realistic effects than in low-dimensional models. Furthermore, even when the sample size is not large enough to use standard full-system VAR techniques, our method remains reliable and delivers accurate results.

Note that unlike Belloni, Chernozhukov, and Hansen (2014a), we do not give a “truly” causal interpretation to the established Granger causalities. In how far Granger causality is a useful concept to study true causality is (and has long been) open to debate, see for example (Eichler, 2013) and the references therein. Moreover, though it appears desirable and in line with Granger's (1969) original intentions to make the information set as large as possible, it is well known in the literature on graphical models (see Eichler, 2013) for causality that considering only the full model is not sufficient for establishing true causal relations from Granger causal ones. For instance, one-period Granger causality in systems with more than two variables cannot capture indirect causal chains spanning over multiple periods. However, the analysis of the full model is a necessary ingredient for any study of causality in a graphical framework. It would therefore be an interesting avenue for further research to study how the method proposed here could fit into such a graphical framework.

Appendix A Proofs

Lemma 2.1. *Let $\boldsymbol{X}_{-GC}$ satisfy Assumption 2(g). Then with probability at least $1-\Delta_T$, we have that*

$$\|\boldsymbol{\delta}\|_2^2 \leq \bar{s}_T \left\| \widetilde{\boldsymbol{X}}^{\otimes}_{-GC} \boldsymbol{\delta}/\sqrt{T} \right\|_2^2 / \phi^2_{T,\min},$$

for any $\boldsymbol{\delta} = (\boldsymbol{\delta}_1', \ldots, \boldsymbol{\delta}_{N_I}')'$ such that $|S_\delta| \leq \bar{s}_T$, where $S_\delta = \bigcup_{i=1}^{N_I} \{m : \delta_{i,m} \neq 0\}$.

Proof of Lemma 2.1. As before, let $\boldsymbol{X}_S$ denote the submatrix containing those columns of $\boldsymbol{X}_{-GC}$ corresponding to the elements in S. It follows from Assumption 2(a) that for any $\boldsymbol{\gamma}$ satisfying $|S_{\boldsymbol{\gamma}}| \leq \bar{s}_T$, we have that $\lambda_{\min}(\boldsymbol{X}_{S_\gamma}' \boldsymbol{X}_{S_\gamma}/T) \geq \phi^2_{T,\min}$. Then, with probability $1-\Delta_T$

we have that

$$
\begin{aligned}
&\min_{\boldsymbol{\delta}:|S_{\boldsymbol{\delta}}|\leq \bar{s}_T} \left\| \widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\boldsymbol{\delta}/\sqrt{T}\right\|_2^2 / \|\boldsymbol{\delta}\|_2^2 \\
&= \min_{\boldsymbol{\delta}:|S_{\boldsymbol{\delta}}|\leq \bar{s}_T} \boldsymbol{\delta}'\left(\boldsymbol{G}_T'\boldsymbol{G}_T \otimes \boldsymbol{X}_{-GC}'\boldsymbol{X}_{-GC}/T\right)\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2^2 \\
&= \min_{|S|\leq \bar{s}_T} \min_{\boldsymbol{x}} \boldsymbol{x}'\left(\boldsymbol{G}_T'\boldsymbol{G}_T \otimes \boldsymbol{X}_S'\boldsymbol{X}_S/T\right)\boldsymbol{x}/\|\boldsymbol{x}\|_2^2 \\
&= \min_{|S|\leq \bar{s}_T} \lambda_{\min}(\boldsymbol{G}_T'\boldsymbol{G}_T \otimes \boldsymbol{X}_S'\boldsymbol{X}_S/T) \\
&= \lambda_{\min}(\boldsymbol{G}_T'\boldsymbol{G}_T) \min_{|S|\leq \bar{s}_T} \lambda_{\min}(\boldsymbol{X}_S'\boldsymbol{X}_S/T) \geq C\phi_{T,\min},
\end{aligned}
$$

as $\lambda_{\min}(\boldsymbol{G}_T'\boldsymbol{G}_T) \geq C > 0$ by assumption. Without loss of generality we may then absorb the constant C into $\phi_{T,\min}$. □

Proof of Theorem 2.1. Our proof follows along the lines of the proof of Theorem 1 and 2 of Belloni, Chernozhukov, and Hansen (2014b), with the main distinction that we consider multiple variables of interest, and multiple "treatments" instead of a single one for each. For this purpose we first define some notation. Let $\boldsymbol{\Gamma} = (\gamma_1, \dots, \gamma_{N_X})$ and $\hat{\boldsymbol{\Gamma}} = (\hat{\gamma}_1, \dots, \hat{\gamma}_{N_X})$, and let $\boldsymbol{\gamma}^{\otimes} = \boldsymbol{I}_{N_I} \otimes \boldsymbol{\Gamma}$. Furthermore, let $\mathcal{P}(\boldsymbol{A}) = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$ denote the projection on the space spanned by $\boldsymbol{A}$ and let $\mathcal{M}(\boldsymbol{A}) = \boldsymbol{I} - \mathcal{P}(\boldsymbol{A})$ denote the corresponding residual-maker. By standard partitioned regression algebra we get

$$
\begin{aligned}
\sqrt{T}(\hat{\boldsymbol{\beta}}^{\text{PDS}}_{GC} - \boldsymbol{\beta}_{GC}) = &\underbrace{\left(\widetilde{\boldsymbol{X}}^{\otimes}_{GC}\mathcal{M}(\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}})\widetilde{\boldsymbol{X}}^{\otimes}_{GC}/T\right)^{-1}}_{\boldsymbol{B}_T^{-1}} \\
&\times \underbrace{\widetilde{\boldsymbol{X}}^{\otimes}_{GC}\mathcal{M}(\widetilde{\boldsymbol{X}}_{\hat{S}^{\otimes}})\left[\widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\boldsymbol{\beta}_{-GC} + \widetilde{\boldsymbol{u}}_I\right]/\sqrt{T}}_{\boldsymbol{a}_T}
\end{aligned}
\tag{2.10}
$$

where $\widetilde{\boldsymbol{X}}^{\otimes} = \boldsymbol{G}_T^{\otimes}\boldsymbol{X}^{\otimes} = \boldsymbol{G}_T \otimes \boldsymbol{X}$. We will now show that $\boldsymbol{a}_T = \widetilde{\boldsymbol{E}}^{\otimes\prime}\widetilde{\boldsymbol{u}}_I/\sqrt{T}$ and $\boldsymbol{B}_T = \widetilde{\boldsymbol{E}}^{\otimes\prime}\widetilde{\boldsymbol{E}}^{\otimes}/T + o_p(1)$. Given these results, the limit distribution then follows directly from Assumption 2(b).

We first consider $\boldsymbol{a}_T$. Note that from (2.7) we have that $\widetilde{\boldsymbol{X}}^{\otimes}_{GC} = \boldsymbol{G}_T \otimes [\boldsymbol{X}_{-GC}\boldsymbol{\Gamma} + \boldsymbol{E}] = \widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\boldsymbol{\Gamma}^{\otimes} + \widetilde{\boldsymbol{E}}^{\otimes}$, and therefore we can write

$$\begin{aligned}\boldsymbol{a}_T = \widetilde{\boldsymbol{E}}^{\otimes\prime}\widetilde{\boldsymbol{u}}_I/\sqrt{T} + \underbrace{\boldsymbol{\Gamma}^{\otimes\prime}\widetilde{\boldsymbol{X}}^{\otimes\prime}_{-GC}\mathcal{M}(\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}})\widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\boldsymbol{\beta}_{-GC}/\sqrt{T}}_{\boldsymbol{a}_{T,1}} \\ + \underbrace{\boldsymbol{\Gamma}^{\otimes\prime}\widetilde{\boldsymbol{X}}^{\otimes\prime}_{-GC}\mathcal{M}(\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}})\widetilde{\boldsymbol{u}}_I/\sqrt{T}}_{\boldsymbol{a}_{T,2}} + \underbrace{\widetilde{\boldsymbol{E}}^{\otimes\prime}\mathcal{M}(\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}})\widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\boldsymbol{\beta}_{-GC}/\sqrt{T}}_{\boldsymbol{a}_{T,3}} \\ - \underbrace{\widetilde{\boldsymbol{E}}^{\otimes\prime}\mathcal{P}(\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}})\widetilde{\boldsymbol{u}}_I/\sqrt{T}}_{\boldsymbol{a}_{T,4}}\end{aligned}$$

We will now show that the terms $\boldsymbol{a}_{T,1}, \dots, \boldsymbol{a}_{T,4}$ vanish. For $\boldsymbol{a}_{T,1}$, note that

$$\|\boldsymbol{a}_{T,1}\|_2 \leq \sqrt{T} \underbrace{\left\|\mathcal{M}(\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}})\widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\boldsymbol{\Gamma}^{\otimes}/\sqrt{T}\right\|_2}_{\|\boldsymbol{A}_{T,1,1}\|_2} \underbrace{\left\|\mathcal{M}(\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}})\widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\boldsymbol{\beta}_{-GC}/\sqrt{T}\right\|_2}_{\|\boldsymbol{a}_{T,1,2}\|_2},$$

where for any matrix $\boldsymbol{M}$, the norm $\|\cdot\|_p$ represents the induced l_p-matrix norm $\|\boldsymbol{M}\|_p = \sup_{\boldsymbol{x}\neq 0} \|\boldsymbol{M}\boldsymbol{x}\|_p/\|\boldsymbol{x}\|_p$. As $\hat{S}_j \subseteq \hat{S}_X$ for all $j = 1, \dots, N_X$ and $\hat{S}^{\otimes}_X \subseteq \hat{S}^{\otimes}$, the space spanned by $\boldsymbol{G}^{\otimes}_T\boldsymbol{X}_{\hat{S}^{\otimes}_X} = \boldsymbol{G}_T \otimes \boldsymbol{x}_{\hat{S}_X}$ is a subspace of the space spanned by $\boldsymbol{G}^{\otimes}_T\boldsymbol{X}^{\otimes}_{\hat{S}^{\otimes}}$, and therefore $\left\|\mathcal{M}\left(\boldsymbol{G}^{\otimes}_T\boldsymbol{X}^{\otimes}_{\hat{S}^{\otimes}}\right)\boldsymbol{y}\right\|_2 \leq \left\|\mathcal{M}\left(\boldsymbol{G}^{\otimes}_T\boldsymbol{X}^{\otimes}_{\hat{S}^{\otimes}_X}\right)\boldsymbol{y}\right\|_2$ for any compatible matrix $\boldsymbol{G}^{\otimes}_T$ and vector $\boldsymbol{y}$. Then, using that

$$\begin{aligned}\mathcal{M}\left(\boldsymbol{G}^{\otimes}_T\boldsymbol{X}^{\otimes}_{\hat{S}^{\otimes}_X}\right) &= \mathcal{M}\left(\boldsymbol{G}_T \otimes \boldsymbol{X}_{\hat{S}_X}\right)\widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\boldsymbol{\Gamma}^{\otimes} \\ &= \mathcal{M}\left(\boldsymbol{G}_T \otimes \boldsymbol{X}_{\hat{S}_X}\right)(\boldsymbol{G}_T \otimes \boldsymbol{X}_{-GC}\boldsymbol{\Gamma}) = \boldsymbol{G}_T \otimes \mathcal{M}\left(\boldsymbol{X}_{\hat{S}_X}\right)\boldsymbol{X}_{-GC}\boldsymbol{\Gamma},\end{aligned} \tag{2.11}$$

we find that

$$\begin{aligned}
\|\boldsymbol{A}_{T,1,1}\|_2 &\leq \left\|\mathcal{M}\left(\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}_X}\right)\widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\boldsymbol{\Gamma}^{\otimes}/\sqrt{T}\right\|_2 \\
&\leq \|\boldsymbol{G}_T\|_2\left\|\mathcal{M}(\boldsymbol{X}_{\hat{S}_X})\boldsymbol{X}_{-GC}\boldsymbol{\Gamma}/\sqrt{T}\right\|_2 \\
&\leq \|\boldsymbol{G}_T\|_2\sum_{j=1}^{N_X}\left\|\mathcal{M}(\boldsymbol{X}_{\hat{S}_X})\boldsymbol{X}_{-GC}\boldsymbol{\gamma}_j/\sqrt{T}\right\|_2 \\
&\leq \|\boldsymbol{G}_T\|_2\sum_{j=1}^{N_X}\left\|\mathcal{M}(\boldsymbol{X}_{\hat{S}_j})\boldsymbol{X}_{-GC}\boldsymbol{\gamma}_j/\sqrt{T}\right\|_2.
\end{aligned}$$

Then,

$$\begin{aligned}
\left\|\mathcal{M}(\boldsymbol{X}_{\hat{S}_j})\boldsymbol{X}_{-GC}\boldsymbol{\gamma}_j/\sqrt{T}\right\|_2 &= \min_{\gamma:\gamma_m=0,m\notin\hat{S}_j}\left\|\boldsymbol{X}_{-GC}\boldsymbol{\gamma}_j-\boldsymbol{X}_{\hat{S}_j}\boldsymbol{\gamma}\right\|_2/\sqrt{T} \\
&\leq \|\boldsymbol{X}_{-GC}(\boldsymbol{\gamma}_j-\hat{\boldsymbol{\gamma}}_j)\|_2/\sqrt{T}, \quad j=0,\ldots,N_X,
\end{aligned} \tag{2.12}$$

as $\hat{S}_j=\{m:\hat{\gamma}_{m,j}\neq 0\}$ and therefore the constraint in the minimization is satisfied. It then follows from Assumption 2(e) that $\|\boldsymbol{A}_{T,1,1}\|_2 \leq N_{GC}\delta_T T^{-1/4}$ with probability $1-\Delta_T$.

For $\boldsymbol{a}_{T,1,2}$, from the definition of the best linear predictor it directly follows that

$$\begin{aligned}
\boldsymbol{\gamma}_0 &= \left(\mathbb{E}\boldsymbol{X}^{\otimes\prime}_{-GC}\boldsymbol{X}^{\otimes}_{-GC}\right)^{-1}\mathbb{E}\boldsymbol{X}'_{-GC}(\boldsymbol{X}^{\otimes}_{GC}\boldsymbol{\beta}_{GC}+\boldsymbol{X}^{\otimes}_{-GC}\boldsymbol{\beta}_{-GC}+\boldsymbol{u}_I) \\
&= \boldsymbol{\Gamma}^{\otimes}\boldsymbol{\beta}_{GC}+\boldsymbol{\beta}_{-GC},
\end{aligned}$$

such that we can substitute $\boldsymbol{\beta}_{-GC}=\boldsymbol{\gamma}_0-\boldsymbol{\Gamma}^{\otimes}\boldsymbol{\beta}_{GC}$ in $\boldsymbol{a}_{T,1,2}$ to find

$$\|\boldsymbol{a}_{T,1,2}\|_2 \leq \left\|\mathcal{M}\left(\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}}\right)\widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\boldsymbol{\gamma}_0/\sqrt{T}\right\|_2+\|\boldsymbol{A}_{T,1,1}\|_2\|\boldsymbol{\beta}_{GC}\|_2,$$

where the negligibility of the second term follows directly from the result above plus Assumption 2(d). As $\hat{S}_0\subseteq\hat{S}^{\otimes}$, the first term can be bounded by

$$\begin{aligned}\left\|\mathcal{M}\left(\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}}\right)\widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\gamma_0/\sqrt{T}\right\|_2 &\leq \left\|\mathcal{M}\left(\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}_0}\right)\widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\gamma_0/\sqrt{T}\right\|_2 \\ &\leq \|\boldsymbol{G}_T\|_2 \left\|\boldsymbol{X}^{\otimes}_{-GC}(\gamma_0 - \hat{\gamma}_0)\right\|_2/\sqrt{T} \leq \sqrt{N_I}\delta_T T^{-1/4},\end{aligned}$$

where we use that, for $\hat{\boldsymbol{\beta}}_{\boldsymbol{G}_T} = \mathcal{M}(\boldsymbol{G}_T\boldsymbol{X})\boldsymbol{G}_T\boldsymbol{y}$,

$$\begin{aligned}\|\mathcal{M}(\boldsymbol{G}_T\boldsymbol{X})\boldsymbol{G}_T\boldsymbol{y}\|_2 = \left\|\boldsymbol{G}_T\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{G}_T}\right)\right\|_2 &\leq \left\|\boldsymbol{G}_T\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{I}}\right)\right\|_2 \\ &\leq \|\boldsymbol{G}_T\|_2 \left\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{I}}\right\|_2.\end{aligned}$$

It then follows directly that $\|\boldsymbol{a}_{T,1}\|_2 = O_p(\delta_T^2) = o_p(1)$.

For $\boldsymbol{a}_{T,2}$, let $\boldsymbol{\gamma}_j$ denote the j-th column of $\boldsymbol{\varGamma}^{\otimes}$ and define the noiseless generalized least squares estimator

$$\check{\gamma}^{\otimes}_{j,S} = \underset{\boldsymbol{\gamma}:\gamma_m = 0, m \notin S}{\arg\min} \left\|\widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\boldsymbol{\gamma}^{\otimes}_j - \widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\boldsymbol{\gamma}\right\|_2^2, \qquad j = 1, \ldots, N_{GC}, \tag{2.13}$$

for any compatible index set S, and let $\check{\boldsymbol{\varGamma}}^{\otimes}_S = \left(\check{\gamma}^{\otimes}_{1,S}, \ldots, \check{\gamma}^{\otimes}_{N_X,S}\right)$, such that $\mathcal{M}\left(\widetilde{\boldsymbol{X}}^{\otimes}_S\right)\widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\boldsymbol{\varGamma}^{\otimes} = \widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\left(\boldsymbol{\varGamma}^{\otimes} - \check{\boldsymbol{\varGamma}}^{\otimes}_S\right)$. Then, with probability $1 - \Delta_T$,

$$\begin{aligned}\|\boldsymbol{a}_{T,2}\|_1 &= \left\|\left(\check{\boldsymbol{\varGamma}}^{\otimes}_{\hat{S}^{\otimes}} - \boldsymbol{\varGamma}^{\otimes}\right)' \widetilde{\boldsymbol{X}}^{\otimes\prime}_{-GC}\widetilde{\boldsymbol{u}}_I/\sqrt{T}\right\|_1 \\ &\overset{(i)}{\leq} \sum_{j=1}^{N_{GC}} \left\|\check{\boldsymbol{\gamma}}^{\otimes}_{j,\hat{S}^{\otimes}} - \boldsymbol{\gamma}^{\otimes}_j\right\|_1 \left\|\widetilde{\boldsymbol{X}}^{\otimes\prime}_{-GC}\widetilde{\boldsymbol{u}}_I/\sqrt{T}\right\|_\infty \\ &\overset{(ii)}{\leq} \gamma_T \sum_{j=1}^{N_{GC}} \left\|\check{\boldsymbol{\gamma}}^{\otimes}_{j,\hat{S}^{\otimes}} - \boldsymbol{\gamma}^{\otimes}_j\right\|_1 \\ &\overset{(iii)}{\leq} \sqrt{\bar{s}_T}\gamma_T \sum_{j=1}^{N_{GC}} \left\|\widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\left(\check{\boldsymbol{\gamma}}^{\otimes}_{j,\hat{S}^{\otimes}} - \boldsymbol{\gamma}^{\otimes}_j\right)/T\right\|_2/\phi_{T,\min}\end{aligned}$$

$$
\begin{aligned}
&\overset{(iv)}{\leq} \sqrt{\bar{s}_T}\gamma_T \sum_{j=1}^{N_{GC}} \left\| \widetilde{\boldsymbol{X}}^{\otimes}_{-GC} \left(\hat{\boldsymbol{\gamma}}^{\otimes}_j - \boldsymbol{\gamma}^{\otimes}_j \right) / T \right\|_2 / \phi_{T,\min} \\
&\overset{(v)}{\leq} \frac{\sqrt{\bar{s}_T}\gamma_T}{\phi_{T,\min}} \|\boldsymbol{G}_T\|_2 N_I \sum_{j=1}^{N_X} \|\boldsymbol{X}_{-GC} \left(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j \right) / T\|_2 \\
&\overset{(vi)}{\leq} N_I^{3/2} N_X \frac{\sqrt{\bar{s}_T}\gamma_T}{\phi_{T,\min}} \delta_T T^{-1/4} \leq \delta_T^2.
\end{aligned}
$$

Here inequality (i) uses that

$$
\|\boldsymbol{A}\boldsymbol{x}\|_1 = \sum_{j=1}^{m} |\boldsymbol{a}_{j\cdot} x_i| \leq \|\boldsymbol{x}\|_\infty \sum_{j=1}^{m} \|\boldsymbol{a}_{j\cdot}\|_1 \tag{2.14}
$$

from the dual norm inequality, where $\boldsymbol{A}$ is a generic $m \times n$ matrix $\boldsymbol{A}$ with j-th row denoted as $\boldsymbol{a}_{j\cdot}$, and a $n \times 1$ vector $\boldsymbol{x}$. Letting $\|\boldsymbol{A}\|_{\max} = \max_{i,j} |a_{ij}|$, Step (ii) follows from the fact that $\left\| \widetilde{\boldsymbol{X}}^{\otimes\prime}_{-GC} \widetilde{\boldsymbol{u}}_I / \sqrt{T} \right\|_\infty = \left\| (\boldsymbol{G}_T^\prime \boldsymbol{G}_T \otimes \boldsymbol{X}^\prime_{-GC}) \boldsymbol{u}_I / \sqrt{T} \right\|_\infty \leq \|\boldsymbol{G}_T^\prime \boldsymbol{G}_T\|_{\max} \left\| \boldsymbol{X}^{\otimes\prime}_{-GC} \boldsymbol{u}_I / \sqrt{T} \right\|_\infty \leq \gamma_T$ by Assumption 2(c), while (iii) follows from bounding the l_1-norm by the l_2-norm and applying Lemma 2.1. (iv) follows from the definition of $\check{\boldsymbol{\gamma}}_{\hat{S}}$ as minimizer of the sum of squares and (v) from the properties of the Kronecker product. Finally (vi) follows from Assumption 2(e).

For $\boldsymbol{a}_{T,3}$, define $\check{\boldsymbol{\gamma}}_{0,S} = \arg\min_{\boldsymbol{\gamma}:\gamma_m=0, m\notin S} \left\| \widetilde{\boldsymbol{X}}^{\otimes}_{-GC} \boldsymbol{\gamma}_0 - \widetilde{\boldsymbol{X}}^{\otimes}_{-GC} \boldsymbol{\gamma} \right\|_2^2$ analogously to (2.13). Then we have with probability $1 - \Delta_T$

$$
\begin{aligned}
\|\boldsymbol{a}_{T,3}\|_1 &\overset{(i)}{\leq} \left\| \widetilde{\boldsymbol{E}}^{\otimes\prime} \mathcal{M}(\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}}) \widetilde{\boldsymbol{X}}^{\otimes}_{-GC} \boldsymbol{\gamma}_0 / \sqrt{T} \right\|_1 + \left\| \widetilde{\boldsymbol{E}}^{\otimes\prime} \mathcal{M}(\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}}) \widetilde{\boldsymbol{X}}^{\otimes}_{-GC} \boldsymbol{\Gamma}^{\otimes} \boldsymbol{\beta}_{GC} / \sqrt{T} \right\|_1 \\
&\overset{(ii)}{\leq} \left\| \widetilde{\boldsymbol{E}}^{\otimes\prime} \widetilde{\boldsymbol{X}}^{\otimes}_{-GC} \left(\check{\boldsymbol{\gamma}}_{0,\hat{S}} - \boldsymbol{\gamma}_0 \right) / \sqrt{T} \right\|_1 + \left\| \widetilde{\boldsymbol{E}}^{\otimes\prime} \widetilde{\boldsymbol{X}}^{\otimes}_{-GC} \left(\check{\boldsymbol{\Gamma}}^{\otimes}_{\hat{S}^{\otimes}} - \boldsymbol{\Gamma}^{\otimes} \right) \boldsymbol{\beta}_{GC} / \sqrt{T} \right\|_1 \\
&\overset{(iii)}{\leq} \left\| \check{\boldsymbol{\gamma}}_{0,\hat{S}} - \boldsymbol{\gamma}_0 \right\|_1 \sum_{j=1}^{N_{GC}} \left\| \widetilde{\boldsymbol{X}}^{\otimes\prime}_{-GC} \widetilde{\boldsymbol{e}}^{\otimes}_j / \sqrt{T} \right\|_\infty
\end{aligned}
$$

$$
\begin{aligned}
&+ \|\boldsymbol{\beta}_{GC}\|_\infty \sum_{j=1}^{N_{GC}} \left\| \widetilde{\boldsymbol{X}}^{\otimes\prime}_{-GC} \widetilde{\boldsymbol{e}}_j^{\otimes} / \sqrt{T} \right\|_\infty \sum_{j=1}^{N_{GC}} \left\| \check{\boldsymbol{\gamma}}^{\otimes}_{j,\hat{S}^{\otimes}} - \boldsymbol{\gamma}_j \right\|_1 \\
&\overset{(iv)}{\leq} \sum_{j=1}^{N_{GC}} \left\| \widetilde{\boldsymbol{X}}^{\otimes\prime}_{-GC} \widetilde{\boldsymbol{e}}_j^{\otimes} / \sqrt{T} \right\|_\infty \left[\left\| \check{\boldsymbol{\gamma}}_{0,\hat{S}} - \boldsymbol{\gamma}_0 \right\|_1 + C N_I^{3/2} N_X \frac{\sqrt{\bar{s}_T}}{\phi_{T,\min}} \delta_T T^{-1/4} \right] \\
&\overset{(v)}{\leq} \sum_{j=1}^{N_{GC}} \left\| \widetilde{\boldsymbol{X}}^{\otimes\prime}_{-GC} \widetilde{\boldsymbol{e}}_j^{\otimes} / \sqrt{T} \right\|_\infty \\
&\left[\frac{\sqrt{\bar{s}_T} \|\boldsymbol{G}_T\|_2}{\phi_{T,\min}} \left\| \boldsymbol{X}^{\otimes}_{-GC} \left(\check{\boldsymbol{\gamma}}_0 - \boldsymbol{\gamma}_0 \right) / \sqrt{T} \right\|_2 + C N_I^{3/2} N_X \frac{\sqrt{\bar{s}_T}}{\phi_{T,\min}} \delta_T T^{-1/4} \right] \\
&\overset{(vi)}{\leq} N_{GC} \gamma_T \left[\frac{\sqrt{\bar{s}_T} \sqrt{N_i}}{\phi_{T,\min}} \delta_T T^{-1/4} + C N_I^{3/2} N_X \frac{\sqrt{\bar{s}_T}}{\phi_{T,\min}} \delta_T T^{-1/4} \right] \\
&\leq C N_X N_I^{3/2} \frac{\sqrt{\bar{s}_T} \gamma_T}{\phi_{T,\min}} \delta_T T^{-1/4} \leq \delta_T^2.
\end{aligned}
$$

Inequality (i) follows from the fact that $\boldsymbol{\beta}_{-GC} = \boldsymbol{\gamma}_0 - \boldsymbol{\Gamma}^{\otimes} \boldsymbol{\beta}_{GC}$, while (ii) follows from the definition of $\tilde{\boldsymbol{\gamma}}_{0,S}$ and (2.13). For the first term in (iii) we use (2.14) whereas for the second term we apply it twice to get

$$
\|\boldsymbol{B}\boldsymbol{A}\boldsymbol{x}\|_1 \leq \|\boldsymbol{x}\|_\infty \sum_{i=1}^{p} \|\boldsymbol{b}_{i\cdot}\boldsymbol{A}\|_1 \leq \|\boldsymbol{x}\|_\infty \sum_{i=1}^{p} \|\boldsymbol{b}_{i\cdot}\|_\infty \sum_{j=1}^{m} \|\boldsymbol{a}_{\cdot j}\|_1 \tag{2.15}
$$

for any $p \times n$ matrix $\boldsymbol{B}$. Step (iv) follows from Assumption 2(d) and the results for $\boldsymbol{a}_{T,2}$, while Step (v) applies the same arguments as used therein to $\left\| \check{\boldsymbol{\gamma}}_{0,\hat{S}} - \boldsymbol{\gamma}_0 \right\|_1$. Finally, Step (vi) follows analogoulsy to Step (ii) for $\boldsymbol{a}_{T,2}$ by noting that $\left\| \widetilde{\boldsymbol{X}}^{\otimes\prime}_{-GC} \widetilde{\boldsymbol{e}}_j^{\otimes} / \sqrt{T} \right\|_\infty \leq \|\boldsymbol{G}_T' \boldsymbol{G}_T\|_{\max} \left\| \boldsymbol{X}'_{-GC} \boldsymbol{e}_j / \sqrt{T} \right\|_\infty \leq \gamma_T$, plus using the bound from Assumption 2(c).

Finally, we consider $\boldsymbol{a}_{T,4}$. We get

$$
\|\boldsymbol{a}_{T,4}\|_1 \overset{(i)}{\leq} \left\| \widetilde{\boldsymbol{X}}^{\otimes\prime}_{\hat{S}^{\otimes}} \widetilde{\boldsymbol{u}}_I / \sqrt{T} \right\|_\infty \sum_{j=1}^{N_{GC}} \left\| \widetilde{\boldsymbol{e}}_j^{\otimes\prime} \widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}} \left(\widetilde{\boldsymbol{X}}^{\otimes\prime}_{\hat{S}^{\otimes}} \widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}} \right)^{-1} \right\|_1
$$

$$\begin{aligned}
&\overset{(ii)}{\leq} \gamma_T\sqrt{\bar{s}_T}\sum_{j=1}^{N_{GC}} \left\| \widetilde{\boldsymbol{e}}_j^{\otimes\prime}\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}} \left(\widetilde{\boldsymbol{X}}^{\otimes\prime}_{\hat{S}^{\otimes}}\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}}\right)^{-1}\right\|_2 \\
&\overset{(iii)}{\leq} \gamma_T\bar{s}_T\left\|\left(\widetilde{\boldsymbol{X}}^{\otimes\prime}_{\hat{S}^{\otimes}}\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}}/T\right)^{-1}\right\|_2\sum_{j=1}^{N_{GC}}\left\|\widetilde{\boldsymbol{X}}^{\otimes\prime}_{\hat{S}^{\otimes}}\widetilde{\boldsymbol{e}}_j^{\otimes}/\sqrt{T}\right\|_\infty/\sqrt{T} \\
&\overset{(iv)}{\leq} N_{GC}\gamma_T^2\bar{s}_T T^{-1/2}/\phi_{T,\min} \leq \delta_T^2,
\end{aligned}$$

where step (i) follows from (2.14). For (ii) we bound the l_1-norm with the l_2-norm, using that $\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}}$ contains a subset of the variables in $\widetilde{\boldsymbol{X}}^{\otimes}_{-GC}$ and therefore $\left\|\widetilde{\boldsymbol{X}}^{\otimes\prime}_{\hat{S}^{\otimes}}\widetilde{\boldsymbol{u}}_I^{\otimes}\right\|_\infty \leq \left\|\widetilde{\boldsymbol{X}}^{\otimes\prime}_{-GC}\widetilde{\boldsymbol{u}}_I^{\otimes}\right\|_\infty$ and apply Assumption 2(c). Step (iii) follows from the Cauchy-Schwarz inequality, bounding the l_2-norm by the l_∞-norm and reasoning as for Step (ii) that $\left\|\widetilde{\boldsymbol{X}}^{\otimes\prime}_{\hat{S}^{\otimes}}\widetilde{\boldsymbol{e}}_j^{\otimes}\right\|_\infty \leq \left\|\widetilde{\boldsymbol{X}}^{\otimes\prime}_{-GC}\widetilde{\boldsymbol{e}}_j^{\otimes}\right\|_\infty$. Finally (iv) follows from Assumption 2(c) and Lemma 2.1.

We next consider $\boldsymbol{B}_T$. Using that $\widetilde{\boldsymbol{X}}^{\otimes}_{GC} = \widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\boldsymbol{\Gamma}^{\otimes} + \widetilde{\boldsymbol{E}}^{\otimes}$, we write

$$\begin{aligned}
\boldsymbol{B}_T = \widetilde{\boldsymbol{E}}^{\otimes\prime}\widetilde{\boldsymbol{E}}^{\otimes}/T &+ \underbrace{\boldsymbol{\Gamma}^{\otimes\prime}\widetilde{\boldsymbol{X}}^{\otimes\prime}_{-GC}\mathcal{M}(\boldsymbol{X}^{\otimes}_{\hat{S}^{\otimes}})\widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\boldsymbol{\Gamma}^{\otimes}/T}_{\boldsymbol{B}_{T,1}} \\
&+ \underbrace{\widetilde{\boldsymbol{\Gamma}}^{\otimes\prime}\widetilde{\boldsymbol{X}}^{\otimes\prime}_{-GC}\mathcal{M}(\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}})\widetilde{\boldsymbol{E}}^{\otimes}/T}_{\boldsymbol{B}_{T,2}} + \underbrace{\widetilde{\boldsymbol{E}}^{\otimes\prime}\mathcal{M}(\boldsymbol{X}^{\otimes}_{\hat{S}^{\otimes}})\widetilde{\boldsymbol{X}}^{\otimes}_{-GC}\boldsymbol{\Gamma}^{\otimes}/T}_{\boldsymbol{B}'_{T,2}} \\
&- \underbrace{\widetilde{\boldsymbol{E}}^{\otimes\prime}\mathcal{P}(\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}})\widetilde{\boldsymbol{E}}^{\otimes}/T}_{\boldsymbol{B}_{T,3}}.
\end{aligned}$$

These terms can be handled as the terms for $\boldsymbol{a}_T$. In particular, with probability $1-\Delta_T$, $\|\boldsymbol{B}_{T,1}\|_2 \leq \|\boldsymbol{A}_{T,1,1}\|_2^2 \leq \delta_T^2 T^{-1/2}$, $\|\boldsymbol{B}_{T,2}\|_2 \leq \delta_T^2 T^{-1/2}$ using the same steps as for $\boldsymbol{a}_{T,2}$, and $\|\boldsymbol{B}_{T,3}\|_2 \leq \delta_T T^{-1/2}$ analogously to $\boldsymbol{a}_{T,4}$.

This shows that $\boldsymbol{a}_T = \widetilde{\boldsymbol{E}}^{\otimes\prime}\widetilde{\boldsymbol{u}}_I/\sqrt{T}$ and $\boldsymbol{B}_T = \widetilde{\boldsymbol{E}}^{\otimes\prime}\widetilde{\boldsymbol{E}}^{\otimes}/T + o_p(1)$. It then follows directly from Assumption 2(b) that

$$\begin{aligned}\sqrt{T}\left(\hat{\boldsymbol{\beta}}^{\text{PDS}}_{GC} - \boldsymbol{\beta}_{GC}\right) &= \left(\boldsymbol{G}_T'\boldsymbol{G}_T \otimes \boldsymbol{E}'\boldsymbol{E}\right)^{-1} \boldsymbol{E}^{\otimes}(\boldsymbol{G}_T'\boldsymbol{G}_T \otimes \boldsymbol{I}_T)\boldsymbol{u}_I + o_p(1) \\ &\xrightarrow{d} \mathcal{N}\left(\mathbf{0}, (\boldsymbol{G}'\boldsymbol{G} \otimes \boldsymbol{\Sigma}_{GC|-GC})^{-1}\boldsymbol{\Omega}_{\boldsymbol{G}}(\boldsymbol{G}'\boldsymbol{G} \otimes \boldsymbol{\Sigma}_{GC|-GC})^{-1}\right).\end{aligned}$$

□

Proof of Theorem 2.2. By partitioned regression algebra, we find that

$$\begin{aligned}LM &= \hat{\boldsymbol{\xi}}^{*\prime}\hat{\boldsymbol{\xi}}^{*} - \hat{\boldsymbol{\nu}}^{*\prime}\hat{\boldsymbol{\nu}}^{*} \\ &= \underbrace{\boldsymbol{y}_I^{*\prime}\mathcal{M}(\boldsymbol{X}^{*\otimes}_{\hat{S}^{\otimes}})\boldsymbol{X}^{*\otimes}_{GC}}_{\boldsymbol{a}_T^{*\prime}} \underbrace{\left[\boldsymbol{X}^{*\otimes\prime}_{GC}\mathcal{M}(\boldsymbol{X}^{*\otimes}_{\hat{S}^{\otimes}})\boldsymbol{X}^{*\otimes}_{GC}\right]^{-1}}_{\boldsymbol{B}_T^{*-1}} \underbrace{\boldsymbol{X}^{*\otimes\prime}_{GC}\mathcal{M}(\boldsymbol{X}^{*\otimes}_{\hat{S}^{\otimes}})\boldsymbol{y}_I^{*}}_{\boldsymbol{a}_T^{*}}.\end{aligned}$$

Note that $\boldsymbol{a}_T^*$ and $\boldsymbol{B}_T^*$ are special cases of their counterparts in the proof of Theorem 2.1 with $\boldsymbol{G}_T = \hat{\boldsymbol{\Sigma}}_{T,I}^{-1/2}$. We now show that this choice of $\boldsymbol{G}_T$ satisfies the conditions of Theorem 2.1. We do this by proving that $\boldsymbol{G}_T$ converges to $\boldsymbol{G} = \boldsymbol{\Sigma}_{u,I}^{-1/2}$, and this satisfies the conditions in the theorem.

Consider one particular element (i,j) of $\hat{\boldsymbol{\Sigma}}_{u,I}$, say $\hat{\sigma}_{u,ij}$. Let $\hat{S}_{0,I_i}$ denote the variables selected in $\hat{S}_0$ corresponding to the equation for variable $\boldsymbol{y}_{I_i}$, where $I = \{I_1, \ldots, I_{N_I}\}$, and let $\hat{S}_i = \left(\bigcup_{j=1}^{N_X} \hat{S}_j\right) \cup \hat{S}_{0,i}$ denote all variables selected that are relevant for $\boldsymbol{y}_{I_i}$. We can then write

$$\begin{aligned}\hat{\boldsymbol{\sigma}}_{u,ij} &= \hat{\boldsymbol{\xi}}_{I_i}'\hat{\boldsymbol{\xi}}_{I_j}/T = \boldsymbol{y}_{I_i}'\mathcal{M}(\boldsymbol{X}_{\hat{S}_i})\mathcal{M}(\boldsymbol{X}_{\hat{S}_j})\boldsymbol{y}_{I_j} \\ &= \boldsymbol{u}_{I_i}'\boldsymbol{u}_{I_j}/T + \underbrace{\boldsymbol{\beta}_{-GC,i}'\boldsymbol{X}_{-GC}'\mathcal{M}(\boldsymbol{X}_{\hat{S}_i})\mathcal{M}(\boldsymbol{X}_{\hat{S}_j})\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC,j}/T}_{d_{T,ij,1}} \\ &\quad - \underbrace{\boldsymbol{u}_{I_i}'\mathcal{M}(\boldsymbol{X}_{\hat{S}_i})\mathcal{M}(\boldsymbol{X}_{\hat{S}_j})\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC,j}/T}_{d_{T,ij,2}}\end{aligned}$$

$$
\begin{aligned}
&\underbrace{-\,\boldsymbol{u}_{I_j}'\mathcal{M}(\boldsymbol{X}_{\hat{S}_j})\mathcal{M}(\boldsymbol{X}_{\hat{S}_i})\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC,i}/T}_{d_{T,ji,2}}\\
&\quad\underbrace{+\,\boldsymbol{u}_{I_i}'\left[\boldsymbol{I}_T-\mathcal{M}(\boldsymbol{X}_{\hat{S}_i})\mathcal{M}(\boldsymbol{X}_{\hat{S}_j})\right]\boldsymbol{u}_{I_j}/T,}_{d_{T,ij,3}}
\end{aligned}
$$

where (under H_0) we write $\boldsymbol{y}_{I_i}=\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC,i}+\boldsymbol{u}_{I_i}$.

We can use the same reasoning as used in the proof of Theorem 2.1 to prove that the terms $d_{T,ij,k}$, $k=1,2,3$ are negligible. Let $\gamma_{0,i}$ denote the sub-vector of $\boldsymbol{\gamma}_0$ corresponding to unit i, and note that under the null hypothesis $\boldsymbol{\gamma}_{0,i}=\boldsymbol{\beta}_{-GC,i}$.
Define $\bar{\boldsymbol{\gamma}}_{i,S}=\arg\min_{\boldsymbol{\gamma}:\gamma_m=0,m\notin S}\|\boldsymbol{X}_{-GC}\gamma_{0,i}-\boldsymbol{X}_{-GC}\boldsymbol{\gamma}\|_2^2$. Then, as $\hat{S}_{0,i}\subseteq\hat{S}_i$, we get that

$$
\begin{aligned}
|d_{T,ij,1}|&\leq\left\|\mathcal{M}(\boldsymbol{X}_{\hat{S}_i})\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC,i}\right\|_2\left\|\mathcal{M}(\boldsymbol{X}_{\hat{S}_j})\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC,j}\right\|_2/T\\
&=\left\|\boldsymbol{X}_{-GC}(\hat{\gamma}_{0,\hat{S}_i}-\boldsymbol{\gamma}_{0,i})/\sqrt{T}\right\|_2\left\|\boldsymbol{X}_{-GC}(\hat{\boldsymbol{\gamma}}_{0,\hat{S}_i}-\boldsymbol{\gamma}_{0,j})/\sqrt{T}\right\|_2\\
&\leq\left\|\boldsymbol{X}_{-GC}(\hat{\boldsymbol{\gamma}}_{0,i}-\boldsymbol{\gamma}_{0,i})/\sqrt{T}\right\|_2\left\|\boldsymbol{X}_{-GC}(\hat{\boldsymbol{\gamma}}_{0,j}-\boldsymbol{\gamma}_{0,j})/\sqrt{T}\right\|_2\leq\delta_T^2T^{-1/2}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
|d_{T,ij,2}|&=\left|\boldsymbol{u}_{I_i}'\mathcal{M}(\boldsymbol{X}_{\hat{S}_i\cup\hat{S}_j})\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC,j}/T\right|=\left|\boldsymbol{u}_{I_i}'\boldsymbol{X}_{-GC}(\hat{\boldsymbol{\gamma}}_{0,\hat{S}_i\cup\hat{S}_j}-\boldsymbol{\gamma}_{0,j})/T\right|\\
&\leq\left\|\boldsymbol{u}_{I_i}\boldsymbol{X}_{-GC}/\sqrt{T}\right\|_\infty\left\|\hat{\gamma}_{0,\hat{S}_i\cup\hat{S}_j}-\boldsymbol{\gamma}_{0,j}\right\|_1/\sqrt{T}\\
&\leq\frac{\sqrt{\bar{s}_T}\gamma_T}{\phi_{T,\min}}\left\|\boldsymbol{X}_{-GC}(\hat{\boldsymbol{\gamma}}_{0,j}-\boldsymbol{\gamma}_{0,j})/\sqrt{T}\right\|_2/\sqrt{T}\leq\frac{\sqrt{\bar{s}_T}\gamma_T}{\phi_{T,\min}}\delta_TT^{-3/4}.
\end{aligned}
$$

Finally,

$$
\begin{aligned}
|d_{T,ij,3}|&=\left|\boldsymbol{u}_{I_i}'\mathcal{P}(\boldsymbol{X}_{\hat{S}_i\cup\hat{S}_j})\boldsymbol{u}_{I_j}/T\right|\\
&\leq\sqrt{\bar{s}_T}\left\|\boldsymbol{X}_{\hat{S}_i\cup\hat{S}_j}'\boldsymbol{u}_{I_j}\right\|_\infty\left\|\boldsymbol{u}_{I_i}'\boldsymbol{X}_{-GC}\left(\boldsymbol{X}_{\hat{S}_i\cup\hat{S}_j}'\boldsymbol{X}_{\hat{S}_i\cup\hat{S}_j}\right)^{-1}\right\|_2/T
\end{aligned}
$$

$$\leq \bar{s}_T \left\| \boldsymbol{X}'_{-GC}\boldsymbol{u}_{I_i}/\sqrt{T} \right\|_\infty \left\| \boldsymbol{X}'_{-GC}\boldsymbol{u}_{I_j}/\sqrt{T} \right\|_\infty \left\| \left(\boldsymbol{X}'_{\hat{S}_i \cup \hat{S}_j} \boldsymbol{X}_{\hat{S}_i \cup \hat{S}_j}/T \right)^{-1} \right\|_2 /T$$
$$\leq \bar{s}_T \gamma_T^2 T^{-1}/\phi_{T,\min},$$

where all bounds hold with probability at least $1 - \Delta_T$. Similarly, by Assumption 2(b) we know that there exist a sequence $\delta_T \to 0$, such that with probability $1 - \Delta_T$, we have that $\left| \boldsymbol{u}'_{I_i}\boldsymbol{u}_{I_j} - \sigma_{u,ij} \right| \leq \delta_T$. As $\boldsymbol{\Sigma}u, I$ only contains a finite number (N_I^2) elements, we may then conclude that with probability at least $1 - \Delta_T$, it holds that $\left\| \hat{\boldsymbol{\Sigma}}_{u,I} - \boldsymbol{\Sigma} \right\|_2 \leq \delta_T$.

We have that $0 < c_L \leq \lambda_{\min}(\boldsymbol{\Sigma}_{u,I}) \leq \lambda_{\max}(\boldsymbol{\Sigma}_{u,I}) \leq c_u < \infty$, where the lower bound follows from Assumption 2(a) and the upper bound from the fact that $\boldsymbol{\Sigma}_{u,I}$ has a finite number of elements. As

$$\boldsymbol{x}'\hat{\boldsymbol{\Sigma}}_{u,I}\boldsymbol{x} \leq \boldsymbol{x}'\boldsymbol{\Sigma}_{u,I}\boldsymbol{x} + \left| \boldsymbol{x}' \left(\hat{\boldsymbol{\Sigma}}_{u,I}\boldsymbol{x} - \boldsymbol{x}'\boldsymbol{\Sigma}_{u,I} \right) \boldsymbol{x} \right|$$
$$\leq \|\boldsymbol{x}\|_2^2 \lambda_{\max}(\boldsymbol{\Sigma}_{u,I}) + \|\boldsymbol{x}\|_2 \left\| \hat{\boldsymbol{\Sigma}}_{u,I} - \boldsymbol{\Sigma}_{u,I} \right\|_2^2,$$
$$\boldsymbol{x}'\hat{\boldsymbol{\Sigma}}_{u,I}\boldsymbol{x} \geq \boldsymbol{x}'\boldsymbol{\Sigma}_{u,I}\boldsymbol{x} - \left| \boldsymbol{x}' \left(\hat{\boldsymbol{\Sigma}}_{u,I}\boldsymbol{x} - \boldsymbol{x}'\boldsymbol{\Sigma}_{u,I} \right) \boldsymbol{x} \right|$$
$$\leq \|\boldsymbol{x}\|_2^2 \lambda_{\min}(\boldsymbol{\Sigma}_{u,I}) + \|\boldsymbol{x}\|_2 \left\| \hat{\boldsymbol{\Sigma}}_{u,I} - \boldsymbol{\Sigma}_{u,I} \right\|_2^2,$$

the established the consistency of $\hat{\boldsymbol{\Sigma}}_{u,I}$ then directly yields that $C_1 - \delta_T \leq \lambda_{\min}(\hat{\boldsymbol{\Sigma}}_{u,I}) \leq \lambda_{\min}(\hat{\boldsymbol{\Sigma}}_{u,I}) \leq C_2 + \delta_T$ with probability $1 - \Delta_T$. It then also follows that with probability $1 - \Delta_T$ we can find a $0 < C_1 \leq C_2 < \infty$ such that $c_1 \leq 1/\lambda_{\max}(\hat{\boldsymbol{\Sigma}}_{u,I}) = \lambda_{\min}(\hat{\boldsymbol{\Sigma}}_{u,I}^{-1}) \leq \lambda_{\max}(\hat{\boldsymbol{\Sigma}}_{u,I}^{-1}) = 1/\lambda_{\min}(\hat{\boldsymbol{\Sigma}}_{u,I}) \leq c_2$, such that the conditions of Theorem 2.1 are satisfied for $\boldsymbol{G}_T = \hat{\boldsymbol{\Sigma}}_{u,I}^{-1/2}$.

With this choice of $\boldsymbol{G}_T$ and $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{u,I} \otimes \boldsymbol{\Sigma}_{GC|-GC}$, we have that $\boldsymbol{\Omega}_{\boldsymbol{G}} = \boldsymbol{I}_{N_X} \otimes \boldsymbol{\Sigma}_{GC|-GC}$. Letting $\boldsymbol{Z}_{N_{GC}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{N_{GC}})$, it then follows that

$$LM = \boldsymbol{a}_T^{*\prime}\boldsymbol{B}_T^{*-1}\boldsymbol{a}_T^* \xrightarrow{d} \boldsymbol{Z}'_{N_{GC}} \left(\boldsymbol{I}_{N_X} \otimes \boldsymbol{\Sigma}_{GC|-GC} \right)^{1/2\prime}$$
$$\times \left(\boldsymbol{I}_{N_X} \otimes \boldsymbol{\Sigma}_{u,I} \otimes \boldsymbol{\Sigma}_{GC|-GC} \right)^{-1} \left(\boldsymbol{I}_{N_X} \otimes \boldsymbol{\Sigma}_{GC|-GC} \right)^{1/2} \boldsymbol{Z}_{N_{GC}}$$

$$= \boldsymbol{Z}'_{N_{GC}} \boldsymbol{Z}_{N_{GC}} = \chi^2_{N_{GC}}. \qquad \square$$

Appendix B Additional Simulation Results

Table 2.5: Simulation results for the bivariate Granger causality test

DGP	Size/Power	ρ	$K\backslash T$	50	100	200	500
2	Size	0	10	5.9	6.6	7.8	11.8
			20	5.6	5.9	7.8	11.8
			50	4.3	7.0	9.7	14.5
			100	5.5	6.7	8.9	13.9

Notes: Size is reported for DGP 2, as described in Section 2.5, for 1000 replications. $T = (50, 100, 200, 500)$ is the time series length, $K = (10, 20, 50, 100)$ the number of variables in the system, the lag length is fixed to $p = 1$. ρ indicates the correlation employed to simulate the time series with the Toeplitz covariance matrix.

Table 2.2: Simulation results for the PDS-LM Granger causality test ($\rho = 0.7$)

DGP	Size/Power	ρ	T	50					100					200					500				
			K	AIC	BIC	EBIC	λ^{th}	λ^{TSCV}	AIC	BIC	EBIC	λ^{th}	λ^{TSCV}	AIC	BIC	EBIC	λ^{th}	λ^{TSCV}	AIC	BIC	EBIC	λ^{th}	λ^{TSCV}
1	Size	0.7	10	5.8	5.1	5.4	6.1	5.7	5.8	5.7	5.2	5.2	5.8	5.1	4.6	4.6	5.0	4.5	4.3	4.0	4.2	4.0	4.2
			20	7.3	6.4	6.1	6.0	7.2	6.0	6.1	6.2	6.3	6.5	5.1	5.5	5.5	5.8	5.9	5.0	4.9	4.9	4.4	5.0
			50	7.6	5.5	6.0	7.5	6.8	6.4	7.3	5.8	7.3	6.7	6.5	4.9	4.2	6.2	5.5	5.7	4.8	4.8	5.2	5.3
			100	6.4	8.1	6.6	NA	8.0	7.7	6.4	6.2	7.1	6.8	6.9	5.8	4.4	5.3	4.6	5.6	3.6	3.6	4.2	3.8
1	Power	0.7	10	19.0	18.2	18.8	20.3	19.5	33.7	34.2	34.3	35.2	33.4	56.6	57.1	56.8	57.8	55.6	93.7	94.2	94.2	94.3	93.3
			20	15.3	18.3	19.1	18.5	17.9	29.2	30.7	31.6	30.7	29.8	54.4	56.8	56.6	56.3	54.8	92.8	94.1	94.1	93.8	93.4
			50	10.3	14.1	19.3	15.2	14.6	24.0	30.4	32.7	31.1	30.3	50.1	57.1	58.8	56.7	51.7	90.7	92.0	92.5	92.1	91.4
			100	9.0	14.0	19.2	NA	15.4	13.6	29.2	33.6	25.4	24.1	34.1	53.6	55.7	49.4	48.9	88.1	94.0	94.5	92.4	91.1
2	Size	0.7	10	6.4	5.8	5.5	5.8	6.8	5.0	5.1	5.3	5.0	5.5	5.2	4.9	5.0	4.5	5.1	4.7	5.0	4.9	4.5	5.1
			20	7.5	6.5	5.6	5.9	5.5	4.8	5.5	5.3	5.8	5.2	5.1	5.0	4.8	5.0	5.8	4.8	4.5	4.9	4.4	6.3
			50	6.8	5.9	6.8	6.0	6.2	7.4	5.3	6.1	5.2	6.3	5.8	5.9	5.7	5.2	6.2	5.6	5.5	5.6	5.9	6.5
			100	8.1	6.5	6.8	6.9	6.5	8.1	5.8	6.2	7.0	5.9	6.7	6.1	6.3	5.5	5.9	5.3	4.4	4.0	4.5	5.0
2	Power	0.7	10	15.4	15.1	16.0	14.5	14.3	26.0	28.1	28.8	26.6	26.5	47.4	49.7	51.6	49.2	51.0	83.2	84.6	85.3	84.6	86.2
			20	12.6	15.8	19.3	15.7	15.5	26.5	27.9	30.0	26.7	26.9	48.4	51.6	53.2	51.1	51.8	83.1	85.3	86.7	85.4	85.9
			50	9.6	15.4	18.4	13.0	12.2	19.1	28.5	31.5	26.1	26.3	40.3	50.5	52.2	47.0	48.9	83.9	87.4	88.8	87.2	87.2
			100	10.3	16.9	21.3	10.7	11.5	12.6	29.3	32.7	21.0	21.4	32.5	53.0	56.0	45.4	49.8	78.7	88.6	89.1	85.4	85.4
3	Size	0.7	10	4.8	4.8	4.7	4.3	5.5	5.3	4.6	4.8	4.6	5.1	5.5	5.5	5.6	5.6	5.4	4.8	4.8	5.0	4.8	5.1
			20	5.6	5.4	5.2	5.1	5.6	5.4	5.8	5.6	5.2	5.2	4.4	4.0	4.0	4.1	4.5	4.5	4.0	4.2	4.2	4.4
			50	8.3	6.0	5.3	5.5	7.0	6.1	5.9	5.2	5.9	6.4	5.6	4.8	4.6	5.2	5.1	6.2	6.8	6.7	6.9	6.7
			100	7.6	6.3	6.3	5.6	6.8	7.2	5.2	5.2	6.1	6.4	5.6	5.8	5.8	6.2	4.7	3.6	4.1	3.7	4.4	4.6
3	Power	0.7	10	9.9	9.8	10.3	10.3	10.7	16.2	16.6	16.4	16.8	16.9	31.5	31.3	31.4	31.6	31.5	68.4	68.9	68.9	68.9	68.8
			20	8.1	8.8	8.2	9.1	9.5	15.7	16.4	16.2	16.4	16.8	29.2	29.0	29.0	28.7	28.3	66.5	68.5	69.4	68.8	67.9
			50	9.7	9.4	10.1	9.7	10.2	15.7	16.1	17.4	16.2	16.6	31.3	31.2	31.7	31.8	29.9	65.1	66.6	67.4	66.9	65.0
			100	8.4	9.4	9.3	9.7	9.9	12.0	15.3	17.7	14.6	15.6	26.4	30.9	31.9	30.2	29.3	66.2	67.8	68.3	68.3	66.5

Notes: Size and Power for the different DGPs described in Section 4.1 are reported for 1000 replications. $T = (50, 100, 200, 500)$ is the time series length, $K = (10, 20, 50, 100)$ the number of variables in the system, the lag-length is fixed to $p = 1$. ρ indicates the correlation employed to simulate the time series with the Toeplitz covariance matrix. The different choices of the tuning parameter λ are reported as: AIC, BIC, EBIC for information criteria, λ^{th} for the theoretical plug-in and TSCV for time series cross-validation as explained in Section 2.5.

Table 2.3: Simulation results for the PDS-WALD Granger causality test

DGP	Size/Power	ρ	T			50					100					200					500		
			K	AIC	BIC	EBIC	λ^{th}	λ^{TSCV}	AIC	BIC	EBIC	λ^{th}	λ^{TSCV}	AIC	BIC	EBIC	λ^{th}	λ^{TSCV}	AIC	BIC	EBIC	λ^{th}	λ^{TSCV}
1	Size	0	10	7.1	6.6	6.0	7.0	7.6	7.0	6.4	6.1	6.8	7.1	5.5	4.3	4.6	4.7	6.9	4.3	4.1	4.3	4.3	4.8
			20	8.0	6.4	6.8	6.5	6.4	6.2	4.9	4.9	5.7	6.9	4.2	4.8	5.5	4.4	5.0	4.3	4.1	4.3	4.1	4.7
			50	7.2	6.2	6.2	7.2	6.2	7.6	6.4	6.6	6.4	7.1	7.1	5.9	5.9	6.3	7.0	6.8	6.5	6.7	6.8	7.1
			100	7.1	7.4	7.3	NA	7.3	8.0	5.2	5.1	6.1	5.9	6.6	4.7	4.9	5.0	5.3	5.8	3.9	4.0	5.2	4.5
1	Power	0	10	31.3	31.8	34.0	31.3	33.8	58.6	59.5	61.2	58.9	58.3	89.1	89.3	89.6	89.3	89.6	99.9	99.9	99.9	99.9	99.9
			20	23.9	27.6	30.9	27.3	25.3	53.4	55.9	58.4	54.6	55.0	85.6	88.2	89.1	86.8	86.1	99.9	100	100	99.9	99.8
			50	15.3	26.0	34.9	19.0	20.4	39.9	54.5	60.0	46.7	46.3	78.8	86.1	87.3	81.4	80.3	99.9	99.9	99.9	99.9	99.9
			100	13.0	21.9	33.1	NA	20.0	21.6	52.6	57.3	31.3	36.7	61.7	85.9	87.3	74.2	74.2	99.5	100	100	99.8	99.7
2	Size	0	10	6.5	6.1	6.5	6.4	6.6	5.1	5.1	5.8	5.1	6.2	5.1	4.8	5.2	4.8	6.3	3.9	3.9	3.9	3.7	4.2
			20	8.4	6.0	6.7	6.5	7.4	5.3	4.8	5.1	5.2	6.0	5.9	6.0	5.8	5.8	6.9	4.7	3.8	4.3	4.6	5.1
			50	8.3	6.2	5.1	6.7	7.6	8.4	6.4	6.7	7.4	7.4	7.6	6.4	6.2	6.8	7.7	6.4	6.6	6.4	6.2	6.4
			100	6.5	7.6	5.9	NA	7.2	7.4	5.3	6.2	5.6	5.2	5.4	5.6	5.9	4.7	5.1	6.1	4.3	5.1	5.0	4.8
2	Power	0	10	18.7	20.2	21.6	18.9	19.2	38.3	39.6	40.9	38.7	40.3	65.2	65.2	67.4	65.0	65.5	97.4	97.4	97.6	97.3	97.6
			20	16.7	20.9	25.4	19.7	19.0	35.8	40.2	44.9	38.1	38.1	64.8	67.5	69.8	66.2	65.5	97.2	97.5	97.4	97.5	97.4
			50	10.1	15.8	22.7	13.8	14.4	25.4	36.8	43.8	33.6	32.1	57.2	66.8	72.1	62.1	61.6	95.0	96.2	96.8	96.1	95.4
			100	10.0	14.6	25.9	NA	11.4	16.6	35.1	46.5	27.3	26.2	45.1	65.3	74.9	57.3	57.9	94.7	97.3	97.7	96.3	96.4
3	Size	0	10	5.5	5.6	6.1	5.8	5.2	6.0	5.1	5.9	5.9	6.3	3.9	4.1	6.1	4.2	4.6	4.1	4.1	3.9	4.0	4.2
			20	4.8	5.5	6.3	5.3	5.5	4.7	4.4	7.4	4.3	4.3	5.4	5.6	9.6	4.7	4.8	4.7	4.4	4.3	4.6	4.8
			50	7.9	6.7	7.8	7.4	7.0	6.7	7.2	9.6	6.1	5.6	7.0	6.9	12.7	6.1	6.8	5.0	5.3	6.6	5.3	5.4
			100	7.4	7.0	8.4	NA	7.7	7.0	6.0	8.6	5.9	6.0	4.7	6.3	10.8	4.2	4.8	4.4	5.1	6.7	5.1	4.8
3	Power	0	10	16.0	20.7	24.5	16.6	17.4	32.1	36.8	44.6	32.9	31.9	58.6	61.4	63.8	59.1	59.6	95.2	95.6	95.7	95.5	95.3
			20	14.3	19.9	27.2	14.2	14.8	29.9	37.8	48.8	30.6	30.3	56.7	62.2	70.0	57.2	55.7	94.1	94.6	94.8	94.5	94.3
			50	12.6	21.7	28.8	13.7	11.8	24.5	40.5	52.8	27.3	27.9	50.6	59.7	73.8	52.5	51.8	91.2	92.8	93.4	92.4	90.9
			100	9.8	20.0	27.7	NA	14.8	15.4	43.0	55.5	20.3	23.0	41.7	62.2	75.2	45.3	45.0	90.0	94.2	95.0	91.5	90.2
1	Size	0.7	10	6.2	5.5	5.9	6.4	5.6	5.9	5.9	5.4	5.4	5.2	5.1	4.7	4.7	5.1	4.4	4.5	4.1	4.2	4.1	4.2
			20	7.8	6.4	6.2	6.4	7.1	6.5	6.2	6.2	6.4	6.3	5.1	5.5	5.5	5.8	5.7	5.1	4.9	4.9	4.6	5.1
			50	9.0	6.4	6.9	8.3	6.4	6.9	7.5	6.1	7.5	6.6	6.6	5.1	4.3	6.5	5.5	5.7	4.8	4.9	5.2	5.3
			100	8.2	8.8	6.9	7.5	8.1	8.3	6.5	6.4	7.4	6.5	7.0	5.8	4.6	5.3	4.6	5.6	3.6	3.7	4.2	4.7
1	Power	0.7	10	19.7	18.9	19.8	21.1	19.3	34.7	35.2	35.2	35.9	33.9	56.7	57.3	57.1	57.8	55.7	93.7	94.2	94.2	94.3	93.9
			20	16.2	19.3	20.0	19.0	17.9	29.7	31.0	32.2	31.5	29.8	54.7	57.1	56.8	56.5	55.3	92.8	94.2	94.2	93.5	93.4
			50	11.3	14.9	20.1	16.4	15.1	24.9	30.8	33.7	31.9	30.8	50.6	57.3	59.0	56.9	52.2	90.7	92.0	92.5	92.1	91.4
			100	9.6	14.9	20.3	12.4	15.4	15.4	29.6	34.0	25.8	24.6	34.5	53.8	56.0	49.7	49.3	88.2	94.1	94.6	92.4	91.2

Notes: Size and Power for the different DGPs described in Section 2.5 are reported for 1000 replications. $T = (50, 100, 200, 500)$ is the time series length, $K = (10, 20, 50, 100)$ the number of variables in the system, the lag-length is fixed to $p = 1$. ρ indicates the correlation employed to simulate the time series with the Toeplitz covariance matrix. NAs are placed whenever the post-OLS estimation was not feasible due to $\hat{s} > T$. The different choices of the tuning parameter λ are reported as: AIC, BIC, EBIC for information criteria, λ^{th} for the theoretical plug-in and TSCV for time series cross-validation as explained in Section 2.5.

Table 2.4: Simulation results for the PDS-LM Granger causality test (Overspecified lag-length)

DGP	Size/Power	ρ	T	50					100					200					500				
			K	AIC	BIC	EBIC	λ^{th}	λ^{TSCV}	AIC	BIC	EBIC	λ^{th}	λ^{TSCV}	AIC	BIC	EBIC	λ^{th}	λ^{TSCV}	AIC	BIC	EBIC	λ^{th}	λ^{TSCV}
1	Size	0	10	7.0	6.4	6.3	6.7	7.2	5.9	6.3	5.8	6.7	5.7	5.6	5.0	5.1	5.2	5.4	5.3	5.4	5.4	5.2	5.6
			20	8.4	7.5	6.7	8.3	7.3	6.3	6.0	5.6	4.9	6.0	5.8	5.6	5.6	5.7	5.7	5.0	3.7	3.7	4.7	4.3
			50	NA	5.4	4.6	NA	NA	8.7	5.0	5.4	6.7	5.9	7.4	5.2	4.9	6.6	6.1	5.7	4.7	5.0	4.7	4.9
			100	NA	NA	5.2	NA	NA	NA	5.0	5.2	NA	NA	7.0	4.3	4.0	5.9	5.5	5.9	5.1	5.3	5.6	4.3
1	Power	0	10	20.2	23.3	24.7	22.5	21.0	45.4	48.7	50.0	47.2	44.8	82.3	83.2	83.8	82.6	81.5	99.7	99.8	99.8	99.8	99.7
			20	13.8	17.4	22.6	16.5	15.1	33.0	44.3	47.5	38.2	38.1	72.9	78.9	79.4	75.2	74.4	99.6	99.6	99.6	99.6	99.6
			50	NA	NA	23.8	NA	NA	16.6	43.0	47.8	29.9	29.6	55.9	77.3	79.8	68.2	66.9	98.0	99.7	99.7	99.0	98.9
			100	NA	NA	23.9	NA	NA	NA	37.2	45.5	NA	NA	20.0	76.3	79.1	53.7	53.1	96.0	99.8	99.8	99.2	94.2
2	Size	0	10	5.4	5.0	4.3	5.2	6.0	6.8	6.9	6.4	6.8	5.9	5.8	5.3	6.0	5.4	5.8	5.1	5.1	5.0	5.1	4.8
			20	7.3	5.6	5.7	7.2	6.6	5.1	4.5	5.0	5.0	5.5	4.9	5.0	5.2	5.6	5.2	4.4	4.4	4.7	4.2	4.1
			50	NA	5.4	5.3	NA	NA	6.4	5.3	6.3	4.7	4.9	8.3	5.2	5.8	5.2	6.1	6.3	5.2	4.8	4.2	4.5
			100	NA	NA	5.7	NA	NA	NA	4.6	5.4	NA	4.6	7.6	5.0	5.9	5.0	6.5	6.1	5.1	4.9	4.5	5.4
2	Power	0	10	13.1	14.7	14.8	14.5	13.6	29.0	29.9	31.3	29.1	28.7	52.7	52.5	55.0	53.5	53.9	94.1	94.3	94.4	94.5	93.8
			20	10.8	15.7	18.7	15.4	13.7	24.0	28.0	34.2	26.8	26.2	52.5	54.9	58.2	54.3	51.7	92.6	92.7	93.0	92.8	91.9
			50	NA	NA	14.9	NA	NA	9.2	26.5	33.7	23.5	20.5	38.4	53.0	61.0	50.0	46.5	88.7	92.1	93.5	91.6	90.3
			100	NA	NA	20.3	NA	NA	NA	26.9	37.9	NA	5.7	13.4	54.8	62.7	42.3	13.6	83.9	93.6	95.2	91.4	89.2
3	Size	0	10	5.6	5.9	7.9	5.8	4.8	6.0	6.4	9.1	6.0	6.1	4.7	5.1	8.2	4.9	4.9	5.1	5.3	4.9	5.1	5.0
			20	6.4	5.5	7.7	6.3	5.8	5.3	5.7	8.2	3.8	4.6	5.8	5.3	9.0	5.1	5.6	4.2	3.5	3.6	3.5	3.7
			50	NA	6.8	7.7	NA	NA	7.9	6.8	11.8	5.9	5.0	6.2	6.2	13.1	5.2	5.6	4.8	4.5	5.5	4.6	5.1
			100	NA	NA	7.7	NA	NA	NA	6.9	11.4	NA	6.3	7.2	6.3	12.2	5.4	5.4	3.7	5.0	5.5	4.3	3.2
3	Power	0	10	10.0	12.2	16.5	9.8	10.1	22.1	24.8	31.7	22.0	21.7	43.2	44.0	50.0	43.8	43.4	87.4	87.4	87.8	87.3	86.8
			20	8.5	13.4	19.9	9.7	9.4	15.5	23.2	37.0	18.4	18.4	39.8	43.4	53.6	41.2	39.9	85.3	86.8	87.0	86.3	84.9
			50	NA	NA	20.5	NA	NA	10.4	26.3	40.1	18.4	17.9	32.2	45.8	59.0	38.3	35.5	78.3	81.6	82.6	79.9	79.4
			100	NA	NA	20.8	NA	NA	NA	26.4	40.2	NA	NA	14.1	46.5	63.8	27.3	30.0	73.0	83.9	85.6	80.9	77.2
1	Size	0.7	10	7.6	6.9	5.9	6.6	6.5	5.5	5.5	5.7	4.7	6.2	4.7	4.6	5.1	4.9	5.4	4.8	4.8	4.8	4.5	4.5
			20	6.9	7.3	7.4	7.7	6.9	6.3	7.1	5.8	7.2	6.0	5.1	4.7	4.9	5.6	5.6	5.4	5.6	5.3	6.0	5.1
			50	NA	NA	7.2	NA	NA	8.6	6.2	5.5	7.2	6.3	7.8	4.6	4.4	5.9	5.0	6.5	5.3	4.9	5.6	5.8
			100	NA	NA	6.5	NA	NA	NA	6.1	5.8	NA	5.2	6.3	5.1	4.5	4.5	5.3	5.9	4.3	4.6	3.9	4.3
1	Power	0.7	10	13.0	13.6	14.4	14.4	14.2	22.7	24.0	25.3	24.7	24.0	42.1	43.4	43.9	43.4	42.7	86.0	87.5	87.5	87.6	86.6
			20	10.2	11.5	12.5	12.8	13.2	19.6	21.1	20.9	21.1	20.4	39.4	42.3	43.2	42.0	40.9	84.8	86.8	86.7	86.4	85.0
			50	NA	NA	15.0	NA	NA	12.6	22.3	23.6	21.6	20.2	30.6	43.6	46.4	41.8	38.8	80.7	85.8	86.0	84.4	82.8
			100	NA	NA	13.2	NA	NA	NA	19.9	23.1	NA	NA	14.3	41.3	44.2	35.0	34.7	72.3	85.2	86.1	82.8	81.5

Notes: Size and Power for the different DGPs described in Section 2.5 are reported for 1000 replications. $T = (50, 100, 200, 500)$ is the time series length, $K = (10, 20, 50, 100)$ the number of variables in the system, the lag-length is fixed to $p = 1$. ρ indicates the correlation employed to simulate the time series with the Toeplitz covariance matrix. NAs are placed whenever the post-OLS estimation was not feasible due to $\hat{s} > T$. The different choices of the tuning parameter λ are reported as: AIC, BIC, EBIC for information criteria, λ^{th} for the theoretical plug-in and TSCV for time series cross-validation as explained in Section 2.5.

Appendix C Additional Material for the Empirical Application

Algorithm 2 Heteroskedasticity-robust PDS-LM Granger causality test

[1] Obtain $\boldsymbol{X}^{*\otimes}$ and $\hat{\boldsymbol{\xi}}^{*}$ as in Algorithm 1, and obtain $\hat{\boldsymbol{E}}^{*\otimes} = \boldsymbol{X}^{*\otimes}_{GC} - \boldsymbol{X}^{*\otimes}_{\hat{S}^{\otimes}}\hat{\boldsymbol{\Gamma}}^{*\otimes}$ as the residuals from the multivariate OLS regression of $\boldsymbol{X}^{*\otimes}_{GC}$ on $\boldsymbol{X}^{*\otimes}_{\hat{S}^{\otimes}}$.

[2] Compute element-wise products $\hat{\boldsymbol{\pi}}_j = \hat{\boldsymbol{e}}^{*\otimes}_j \odot \hat{\boldsymbol{\xi}}^{*}$ for $j = 1, \dots, N_{GC}$. Regress a vector of ones on $\hat{\boldsymbol{\Pi}} = (\hat{\boldsymbol{\pi}}_1, \dots, \hat{\boldsymbol{\pi}}_{N_{GC}})$ and compute $TN_I R^2$ from this regression.

[3] Reject H_0 if $TN_I R^2 > q_{\chi^2_{N_{GC}}}(1-\alpha)$, where $q_{\chi^2_{N_{GC}}}(1-\alpha)$ is the $1-\alpha$ quantile of the χ^2 distribution with N_{GC} degrees of freedom.

Table 2.6: Stocks used in Section 2.6

N.	Symbol	Issue name	N.	Symbol	Issue name
1	AAPL	APPLE INC	16	KO	COCA-COLA CO
2	AXP	AMERICAN EXPRESS CO	17	MCD	MCDONALD'S CORP
3	BA	BOEING CO	18	MMM	3M
4	CAT	CATERPILLAR	19	MRK	MERCK & CO
5	CSCO	CISCO SYSTEMS	20	MSFT	MICROSOFT CORPORATION
6	CVX	CHEVRON CORP	21	NKE	NIKE INC
7	DD	DOW CHEMICAL COMPANY	22	PFE	PFIZER INC
8	DIS	WALT DISNEY CO	23	PG	PROCTER & GAMBLE CO
9	GE	GENERAL ELEC	24	TRV	TRAVELERS COMPANIES INC
10	GS	GOLDMAN SACHS GROUP INC	25	UNH	UNITEDHEALTH GROUP INC
11	HD	HOME DEPOT INC	26	UTX	UNITED TECHNOLOGIES CORPORATION
12	IBM	INTL BUS MACHINE	27	V	VISA INC
13	INTC	INTEL CORP	28	VZ	VERIZON COMMUNICATIONS INC
14	JNJ	JOHNSON &JOHNSON	29	WMT	WALMART INC
15	JPM	JPMORGAN CHASE & CO	30	XOM	EXXON MOBIL CORPORATION

3
Inference in Non-stationary High-Dimensional VARs[1]

[1] This chapter is based on a joint work with Alain Hecq and Stephan Smeekes from Maastricht University.

Abstract

In this chapter we use the lag-augmentation idea of Toda and Yamamoto (1995) to build an inferential procedure which holds for high-dimensional unit-root non-stationary VAR models. We prove that we can restrict the augmentation to only the variables of interest for the testing, thereby reducing the loss of power coming from the misspecification of the model. By means of a post-double selection procedure where we use the lasso to reduce the parameter space, we are able to partial-out the effect of nuisance parameters and establish uniform asymptotics. We apply our procedure to the untransformed FRED-MD dataset to investigate the main macroeconomic drivers of inflation.

3.1 Introduction

Learning causes and effects in time series models is a well studied problem in a vast literature going all the way back to the seminal work of Granger (1969). Statistically assessing the predictability among two (or blocks of) time series is to these days a fundamental concept in modern time series analysis. Its applications range over from macroeconomics, finance, network theory, climate econometrics and even the neuroscience. Among others, in the macroeconomic literature, the question of causality between money and gross domestic product (GDP), initiated by Sims et al. (1990) and Stock and Watson (1989a), is still at debate these days, see e.g. Miao et al. (2020). Financial applications of linear and non-linear Granger causality see, among others: Hiemstra and Jones (1994) who find Dow Jones stock returns and percentage changes in New York Stock Exchange trading volume to be bi-directional Granger causing; Billio, Getmansky, et al. (2012) which in a network framework uses principal component techniques as well as Granger causality networks to measures the connectedness among monthly returns of hedge funds, banks, broker/dealers, and insurance companies. Many applications are also found in climate science, for instance in trying to understand and disentangle the causes of climate change. Among others, Stern and Kaufmann (2014) investigate causality between greenhouse gases transformed into radiative forcing and temperature, finding that both natural and anthropogenic forcings cause an upward temperature change and that temperature causes greenhouse gas concentration changes. In neuroscience, Granger causality is widely employed in understanding principles and mechanisms underlying complex brain function and cognition. Examples lies mostly in the branch of functional neuroimaging, where brain connectivity is investigated through neuronal networks from fMRI, EEG, and electrocorticography data (see e.g. Seth et al. (2015), Friston et al. (2013) for some reviews).

More recently, with the increased availability of larger datasets, these causality concepts have been extended to a high-dimensional setting where they can benefit from the inclusion of many more series within the available information set. Granger causality (in mean), as conceived by Granger himself (Granger, 1969) is in fact well known to capture predictability among variables (or blocks of variables) of interest, conditionally on a given information set. In other words, to talk about true direct causal effects, the tested relation must be conditioned upon all possible variables that can aid in explaining the original variables object of the test. Otherwise, omitted variable bias would invalidate the causal interpretation (spurious causality) and the testing procedure would reduce to a mere predictability exercise. Granger himself envisioned this information set as "all the knowledge in the universe available at that time" (Granger, 1980, p.330). As this concept can be hardly operationalized, one needs to rely on the available dataset. In this sense, the high dimensionality of the nowadays increasingly large datasets available, as well as the regularization techniques developed to circumvent the curse of dimensionality and simultaneously perform variable selection and parameter estimation, provide a great opportunity to get-away from spurious causality and approach the true causal interpretation as envisioned by Granger. In Chapter 2 we designed a Granger causality Lagrange-Multipliers (LM) test for high-dimensional vector autoregressive models (VAR) which combines dimensionality reduction techniques based on penalized regressions such as the lasso of Tibshirani (1996), with the post-double selection procedure of Belloni, Chernozhukov, and Hansen (2014b) designed to guarantee uniform asymptotic validity of the post-selection least squares estimator. Empirical applications of such testing procedure comprise, among others, networks construction for volatility spillovers which can be used to predict the flow of volatility contagion when a financial crisis hits the stock market.

In Chapter 2 we assumed stationarity of all the time series considered. This is a long standing issue in econometrics: on the one hand, working with stationary time series alleviates many complications in the asymptotic analysis, allowing for standard inferential procedures such as χ^2 and F-tests. On the other hand, it assumes prior knowledge of the integration and cointegration order and possibly of the type of non-stationarity for all the time series entering the model. This prior knowledge is usually acquired via unit root and cointegration tests such as e.g., the Augmented Dicky-Fuller test (ADF) (see Dickey and Fuller, 1979) and Johansen's cointegration test (see Johansen, 1991). However, these tests are particularly keen to biases coming from different sources. For instance, the inclusion or not of the intercept and the deterministic time trend in the regression equation, the choice of the lag-length order to include, the seasonality adjustments of the data, the presence of structural breaks as well as outliers in the series, these are all factors affecting the outcomes of these tests. Especially unit root tests are known to suffer from low power (see e.g. the critique of Cochrane, 1991). Given that the practitioner should test for unit roots all the time series in a high-dimensional VAR, it follows that biases would accumulate quite dramatically. Furthermore, in practice many observed time series in e.g., macroeconomics, finance, climate econometrics, they definitely do not appear to be stationary in their original levels but they are characterized by stochastic and/or deterministic trends. Taking the first differences of the series (difference-stationary) stabilises the mean by removing changes in the levels and thereby eliminating (or reducing) trend and seasonality. However, this is often not an innocuous transformation: it can indeed induce a loss of information since the long-term memory of the series gets wiped out by the differentiation. Also, when estimating with least squares difference-stationary series that are truly cointegrated, the model gets misspecified. Vector error-correction models (VECM) account for the latter issue since they allow for both short and long memory dynamic in the relationship. However, to be able to write the VECM one typically again relies on unit root and cointegration pre-tests (notable exception is the work of Smeekes and Wijler, 2021).

The aim of this chapter is to design a method which allows for testing Granger causality in high-dimensional VAR models, irrespectively of the integration and cointegration orders of the time series entering the VAR equations. Namely, we seek to avoid any unit root and cointegration pre-test biases and use the VAR in levels directly to perform inference on the parameters of interest. To do so, we borrow the idea of Toda and Yamamoto (1995) which consider a simple lag-augmentation of the system in order to reconduct the asymptotics to standard stationary arguments. We find that the potential inefficiency of this method, coming from the overspecification of the model lag-length, is substantially reduced if the augmentation is performed only on the interest variable(s) for the Granger causality testing. Put it differently, we do not need to augment lags of all the variables within the information set but only the lags of those variables that we are interested in testing for causality. Using the same notation introduced in Chapter 2, let $N_I = |I|$ and $N_J = |N_J|$ denote the number of variables in I and J i.e., respectively the set of Granger caused and Granger causing variables. Then, the lag augmentation can be confined to only $N_I + N_J$ variables. This applies as long as $N_I + N_J$ is sufficiently small. We argue that the relative small dimension of the causal blocks, is not a restriction. After all, the value added of the high-dimensionality approach is that of being able to condition simple relations among series (i.e., bivariate/trivariate), to a large set of regressors, thus in order to maximise the information set to obtain a result as much as possible free of omitted variable bias. To account for the large dimensionality of the VAR and in order to deal with the complications of the post-selection inference (see e.g. Leeb and Pötscher, 2005), we follow the framework outlined in Chapter 2, extending the stationary setting to the unit root non-stationary one. Hence, we set up a post-double selection procedure which is able to partial-out nuisance variables while safeguarding from omitted variable bias to return uniform asymptotics. After the selection has been performed, we build a post-selection, restricted-lag-augmented Lagrange Multiplier test which allows to perform inference on the interest parameters. Several technical assumptions are needed to extend the post-double selection framework to the non-stationary setting. Especially, in order to bound the empirical process, a novel Gaussian approximation is proposed which avoids assuming any strong invariance principles.

The remainder of the chapter is organized as follows: Section 3.2 introduces the model, the relevant hypotheses and the algebra required to show how a lag-augmentation, restricted to the sole variables of interest, does not impact the tested hypothesis which can equivalently be rewritten in terms of the augmented model. Section 3.3 shows how the post-double selection technique can be coupled with a post-selection restricted lag-augmentation as advocated in the previous section. The main algorithm (PDS-LA-LM) is stated and some remarks are given. The set of assumptions for the asymptotic normality of the post-selection least square estimator are extended to the unit root non-stationary framework in Section 3.4 where the main asymptotic results are stated, namely in high-dimensions the restricted lag-augmentation does return asymptotic normality of the OLS after the double-selection is performed by the lasso and hence standard inference, free of omitted variable bias, is attained. In Section 3.5 we report the finite sample simulations for sparse and non-sparse data generating processes and we discuss the performances of the proposed test. In Section 3.6 we elaborate on how to obtain an empirical upper-bound for the estimated lag-length p by means of using information criteria (BIC). The original model is estimated as a diagonal VAR and an estimate of the log-determinant of the residuals covariance matrix is used to alleviates singularity issues affecting the computation of the BIC. Section 3.7 uses the proposed testing framework to investigate the driving factors of inflation in the context of the FRED-MD dataset. Finally, Section 3.8 concludes.
The Appendices are organized as follows. In Appendix A are reported complementary results to the following two appendices. Appendix B reports the proofs of theorems and lemmas needed for the results stated in Section 3.4. The proofs of the main results on the post-double selection estimator convergence is reported in Appendix C. Appendix D reports additional simulation results.

A few words on notation. For any n-dimensional vector $\boldsymbol{x}$, we let $\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ denote the ℓ_p-norm. For any index set $S \subseteq \{1, \dots, n\}$, let $\boldsymbol{x}_S$ denote the sub-vector of $\boldsymbol{x}_t$ containing only those elements x_i

such that $i \in S$. $|S|$ denotes the cardinality of the set S. We use $\xrightarrow{p}$ and $\xrightarrow{d}$ to denote convergence in probability and distribution, respectively.

3.2 The Model

Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$ be a K-dimensional multiple time series process, where $\boldsymbol{y}_t = (y_{1,t}, \ldots, y_{K,t})'$ is generated by a VAR(p) process

$$\boldsymbol{y}_t = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t, \qquad t = p+1, \ldots, T \ , \tag{3.1}$$

where $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p$ are $K \times K$ parameter matrices and $\boldsymbol{u}_t$ is a $K \times 1$ martingale difference sequence (mds) of error terms. Elements of $\boldsymbol{y}_t$ are time series integrated of order d: $I(d)$ $(d = 0, 1, 2)$ and possibly cointegrated of order d, b: $CI(d, b)$.

Assumption 3. The VAR model in (3.1) satisfies:

(a) $\{\boldsymbol{u}_t\}_{t=1}^T$ is a mds with respect to $\mathcal{F}_t = \sigma(\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, \ldots)$ $\boldsymbol{u}_t$ such that $\mathbb{E}(\boldsymbol{u}_t|\mathcal{F}_{t-1}) = \boldsymbol{0}$ for all t; the $K \times K$ covariance matrix $\boldsymbol{\Sigma}_u = \mathbb{E}(\boldsymbol{u}_t \boldsymbol{u}_t')$ is positive definite and $\mathbb{E}|\boldsymbol{u}_t|^{2+\delta} \leq \infty$, with $\delta > 0$.

(b) Roots of $\det(\boldsymbol{I}_K - \sum_{j=1}^p \boldsymbol{A}_j z^j)$ can either lie on the unit disc or outside, thus allowing for unit roots and cointegration within the VAR.

Remark 3.1. Assumption 5-8 from Johansen (1992) are also needed to rule out explosive processes and guaranteeing the series to be maximum $I(2)$ and in general cointegrated. The statements of these assumptions are reported in Appendix A.

We are interested in testing the null hypothesis of Granger non-causality in mean between variables in the set J i.e, the Granger causing variables and those in the set I i.e., the Granger caused, conditional on all the other variables, where $J, I \subset \{1, \ldots, K\}$ and $J \cap I = \emptyset$. Let $N_J = |J|$

and $N_I = |I|$ be the cardinalities of the sets J, I. For the moment we assume the lag-length p in (3.1) to be fixed, we shall further elaborate on the choice of p in Section 3.6. Also, in order to ease the notation, we omit both the intercept and any polynomial time trend from the model, the results we derive easily extends to those cases as well. As in Chapter 2, we describe our procedure in general form for testing blocks of variables. For a formal definition of Granger causality we refer to equation (2.2) in Chapter 2. A test for Granger causality is built via testing the joint significance of the blocks of coefficients in the matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p$ corresponding to the impact of variables J on I. We use a similar stacked representation to that in (3.1) of Chapter 2 for the variables in I. Namely, $\boldsymbol{Y} = (\boldsymbol{y}_{p+1}, \ldots, \boldsymbol{y}_T)'$ and $\boldsymbol{y}_I = \text{vec}\,(\boldsymbol{Y}_I)$ denotes the $N_I \times 1$ stacked vector containing all observations corresponding to the variables in I. Similarly, $\boldsymbol{u}_I = \text{vec}(\boldsymbol{U}_I)$, where $\boldsymbol{U} = (\boldsymbol{u}_{p+1}, \ldots, \boldsymbol{u}_T)'$. Let $\boldsymbol{X} = \left(\boldsymbol{x}_p^{(p)}, \ldots, \boldsymbol{x}_{T-1}^{(p)}\right)'$ and $\boldsymbol{X}^{\otimes} = \boldsymbol{I}_{N_I} \otimes \boldsymbol{X}$, while the stacked parameter vector is $\boldsymbol{\beta} = \text{vec}((\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p)')$. Then we obtain

$$\boldsymbol{y}_I = \boldsymbol{X}^{\otimes}\boldsymbol{\beta} + \boldsymbol{u}_I = \boldsymbol{X}^{\otimes}_{\underline{GC}}\boldsymbol{\beta}_{\underline{GC}} + \boldsymbol{X}^{\otimes}_{-GC}\boldsymbol{\beta}_{-GC} + \boldsymbol{u}_I, \tag{3.2}$$

where $\boldsymbol{X}^{\otimes}_{\underline{GC}} = \boldsymbol{I}_{N_I} \otimes \boldsymbol{X}_{\underline{GC}}$, and $\boldsymbol{X}_{\underline{GC}} = \left(\boldsymbol{x}_{J,p}^{(p)}, \boldsymbol{x}_{I,p}^{(p)} \ldots, \boldsymbol{x}_{J,T-1}^{(p)}, \boldsymbol{x}_{I,T-1}^{(p)}\right)'$ contains those columns of $\boldsymbol{X}$ corresponding to the potentially Granger causing variables in J and those potentially Granger caused variables in I;[2] $\boldsymbol{X}_{-GC}$ and $\boldsymbol{X}^{\otimes}_{-GC}$ are then defined similarly but containing the remaining variables.[3] $\boldsymbol{\beta}_{-GC}$ has $(K - N_J - N_I) \times N_I \times p$ elements and $\boldsymbol{\beta}_{\underline{GC}}$ has $N_{GC} = (N_J + N_I) \times N_I \times p$ elements. Elements of $\boldsymbol{\beta}_{-GC}$ are assumed being large given a large number of variables K is assumed. Throughout the chapter we assume N_J, N_I and p to be fixed when sample size T increases to infinity. Similarly, elements of $\boldsymbol{\beta}_{\underline{GC}}$ are also implied to be fixed. Similarly to Chapter 2, let also $\boldsymbol{\beta}_{GC}$ be the subvec-

[2] Note: the underlined notation is used to distinguish the notation from Chapter 2 where $\boldsymbol{X}_{GC}$ was referring to only the Granger causing instead of both Granger causing and Granger caused.

[3] Note that if $I = \{i\}$ for one particular value of interest, then (3.2) simply corresponds to a single equation from the VAR in (3.1).

tor of $\boldsymbol{\beta}_{\underline{GC}}$ corresponding to the variables in $\boldsymbol{X}_{GC} = \left(\boldsymbol{x}_{J,p}^{(p)}, \dots, \boldsymbol{x}_{J,T-1}^{(p)}\right)'$ i.e., pertaining those columns of $\boldsymbol{X}$ corresponding to only the potentially Granger causing variables in J, thus containing $N_J \times N_I \times p$ elements.

Testing for no Granger causality is then equivalent to testing the following null hypothesis:

$$H_0 : \boldsymbol{\beta}_{GC} = \mathbf{0} \quad \text{against} \quad H_1 : \boldsymbol{\beta}_{GC} \neq \mathbf{0}. \tag{3.3}$$

In order to test the null hypothesis in (3.3) we augment the level of the system in the following way:

$$\boldsymbol{y}_I = \boldsymbol{X}_{\underline{GC}}^{\otimes *} \boldsymbol{\beta}_{\underline{GC}}^{*} + \boldsymbol{X}_{-GC}^{\otimes} \boldsymbol{\beta}_{-GC} + \boldsymbol{u}_I, \tag{3.4}$$

where now $\boldsymbol{X}_{\underline{GC}}^{\otimes *} = \boldsymbol{I}_{N_I} \otimes \boldsymbol{X}_{\underline{GC}}^{*}$, and
$\boldsymbol{X}_{\underline{GC}}^{*} = \left(\boldsymbol{x}_{J,p}^{(p+d)}, \boldsymbol{x}_{I,p}^{(p+d)}, \dots, \boldsymbol{x}_{J,T-1}^{(p+d)}, \boldsymbol{x}_{I,T-1}^{(p+d)}\right)'$ contains the same elements of $\boldsymbol{X}_{\underline{GC}}$ plus additional d lags of both variables in J and I. Similarly as above, considering $\boldsymbol{\beta}_{GC}^{*}$ being the subvector of $\boldsymbol{\beta}_{\underline{GC}}^{*}$ only containing coefficients related to variables in $\boldsymbol{X}_{GC} = \left(\boldsymbol{x}_{J,p}^{(p)}, \dots, \boldsymbol{x}_{J,T-1}^{(p)}\right)'$, then the null hypothesis under test of Granger causality in this lag-augmented set up becomes:

$$H_0 : \quad \boldsymbol{\beta}_{GC}^{*} = 0 \qquad \text{against} \qquad H_1 : \quad \boldsymbol{\beta}_{GC}^{*} \neq 0. \tag{3.5}$$

Let us introduce the following $N_{GC}(p+d) \times N_{GC}(p+d)$ transformation matrices $\boldsymbol{R}_d$ for $d = 1, 2$

$$\boldsymbol{R}_1 = \begin{bmatrix} 1 & 0 & 1 & 0 & \dots & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & \dots & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & \dots & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 & 1 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{R}_2 = \begin{bmatrix} 1 & 0 & 1 & 0 & \dots & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & \dots & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & \dots & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Note we can rewrite $\boldsymbol{R}_1,\ \boldsymbol{R}_2$ as block matrices like

$$\boldsymbol{R}_1 = \begin{pmatrix} \underbrace{\boldsymbol{R}_{11}}_{N_{GC}p\times N_{GC}p} & \underbrace{\boldsymbol{R}_{12}}_{N_{GC}p\times N_{GC}d} \\ \underbrace{\boldsymbol{0}}_{N_{GC}d\times N_{GC}p} & \underbrace{\boldsymbol{R}_{22}}_{N_{GC}d\times N_{GC}d} \end{pmatrix}, \quad \boldsymbol{R}_2 = \begin{pmatrix} \underbrace{\boldsymbol{R}_{11}}_{N_{GC}p\times N_{GC}p} & \underbrace{[\ \underbrace{\boldsymbol{R}_{12}}_{N_{GC}p\times d} \mid \underbrace{\boldsymbol{0}}_{N_{GC}p\times d}\ ]}_{N_{GC}p\times N_{GC}d} \\ \underbrace{\boldsymbol{0}}_{N_{GC}d\times N_{GC}p} & \underbrace{\boldsymbol{I}}_{N_{GC}d\times N_{GC}d} \end{pmatrix},$$

where $\boldsymbol{R}_{11}, \boldsymbol{R}_{22}$ are smaller versions of $\boldsymbol{R}_1$ while $\boldsymbol{R}_{12}$ is everywhere as the upper triangle of $\boldsymbol{R}_1$.

Then let

$$\boldsymbol{P}_1 = \boldsymbol{R}_1; \quad \boldsymbol{P}_1^{-1} = \begin{bmatrix} 1 & 0 & -1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{3.6}$$

and

$$\boldsymbol{P}_2 = \boldsymbol{R}_1\boldsymbol{R}_2 = \begin{bmatrix} 1 & 0 & 2 & 0 & 3 & \cdots & p+1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 & 0 & \cdots & 0 & p+1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 2 & \cdots & p & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 & p & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{3.7}$$

$$
\boldsymbol{P}_2^{-1} = \begin{bmatrix}
1 & 0 & -2 & 0 & 1 & \cdots & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & -2 & 0 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & -2 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & 0 & \cdots & 1 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1
\end{bmatrix}.
$$

For the order of integration $d \leq 2$, define[4]

$$
\begin{aligned}
\begin{pmatrix} \boldsymbol{\beta}^*_{\underline{GC},d} \\ \boldsymbol{\beta}_{-GC} \end{pmatrix} &:= \begin{pmatrix} \boldsymbol{P}_d & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{p(K-N_{GC})} \end{pmatrix}' \begin{pmatrix} \boldsymbol{\beta}^*_{\underline{GC}} \\ \boldsymbol{\beta}_{-GC} \end{pmatrix}, \\
\begin{pmatrix} \boldsymbol{X}^{\otimes *}_{\underline{GC},d} \\ \boldsymbol{X}^{\otimes}_{-GC} \end{pmatrix} &:= \begin{pmatrix} \boldsymbol{P}_d & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{p(K-N_{GC})} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}^{\otimes *}_{\underline{GC}} \\ \boldsymbol{X}^{\otimes}_{-GC} \end{pmatrix}.
\end{aligned}
\tag{3.8}
$$

Therefore, we can rewrite (3.4) as

$$
\begin{aligned}
\boldsymbol{y}_I &= \boldsymbol{X}^{\otimes *}_{\underline{GC}} \boldsymbol{P}_d^{-1} \boldsymbol{P}_d \boldsymbol{\beta}^*_{\underline{GC}} + \boldsymbol{X}^{\otimes}_{-GC} \boldsymbol{\beta}_{-GC} + \boldsymbol{u}_I, = \\
&= \boldsymbol{X}^{\otimes *}_{\underline{GC},d} \boldsymbol{\beta}^*_{\underline{GC},d} + \boldsymbol{X}^{\otimes}_{-GC} \boldsymbol{\beta}_{-GC} + \boldsymbol{u}_I \\
&= \boldsymbol{W}^*_d \boldsymbol{\phi}^* + \boldsymbol{X}^{\otimes}_{-GC} \boldsymbol{\beta}_{-GC} + \boldsymbol{u}_I,
\end{aligned}
\tag{3.9}
$$

where to lighten the notation we defined $\boldsymbol{W}^*_d := \boldsymbol{X}^{\otimes *}_{\underline{GC},d} = \boldsymbol{I}_I \otimes \boldsymbol{X}^*_{\underline{GC},d}$ and $\boldsymbol{\phi}^* = \boldsymbol{\beta}^*_{\underline{GC},d}$. Let further $\boldsymbol{w}^*_t$ be the t-th row of $\boldsymbol{X}^{\otimes *}_{\underline{GC}}$ and $\boldsymbol{w}_{t,d}$ be the t-th row of $\boldsymbol{X}^*_{\underline{GC},d}$ and suppose for simplicity that $N_J = N_I = 1$, where the interest is in testing Granger causality from y_2 to y_1, then

[4] We could allow for more generality than $d \leq 2$. However, as it is consensus in economics, rarely processes exhibit roots of higher order than two. Hence, we confine our presentation up to the case of $I(2)$ series.

we have

$$\begin{aligned}\boldsymbol{w}_{t,d} = (&\underbrace{\Delta^d y_{1,t-1}, \Delta^d y_{2,t-1}}_{\Delta^d \boldsymbol{w_{t,1}}}, \underbrace{\Delta^d y_{1,t-2}, \Delta^d y_{2,t-2}}_{\Delta^d \boldsymbol{w_{t,2}}}, \dots \\ &\dots, \underbrace{\Delta^{d-1} y_{1,t-p-1}, \Delta^{d-1} y_{2,t-p-1}}_{\Delta^{d-1} \boldsymbol{w_{t,p+1}}}, \underbrace{y_{1,t-p-2}, y_{2,t-p-2}}_{\boldsymbol{w_{t,p+2}}})',\end{aligned}$$

Example 3.1. Let $N_J = N_I = p = d = 1$ and the interest is in testing Granger causality from $y_{2,t}$ to $y_{1,t}$ then,

$$\underbrace{\boldsymbol{R}_1}_{4\times 4} = \begin{pmatrix} \underbrace{\boldsymbol{R}_{11}}_{2\times 2} & \underbrace{\boldsymbol{R}_{12}}_{2\times 2} \\ \underbrace{\boldsymbol{0}}_{2\times 2} & \underbrace{\boldsymbol{R}_{22}}_{2\times 2} \end{pmatrix}; \ \boldsymbol{P}_d = \boldsymbol{P}_1 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \boldsymbol{R}_1; \ \boldsymbol{P}_1^{-1} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix};$$

$$\boldsymbol{P}_1^{-1}\boldsymbol{w}_t^* = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} = \begin{bmatrix} \Delta y_{1,t-1} \\ \Delta y_{2,t-1} \\ y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} = \boldsymbol{w}_{t,1}.$$

Example 3.2. Let $N_J = N_I = 1$, $p = d = 2$ and the interest is in testing Granger causality from $y_{2,t}$ to $y_{1,t}$ then, then

$$\underbrace{\boldsymbol{R}_1}_{8\times 8} = \begin{pmatrix} \underbrace{\boldsymbol{R}_{11}}_{4\times 4} & \underbrace{\boldsymbol{R}_{12}}_{4\times 4} \\ \underbrace{\boldsymbol{0}}_{4\times 4} & \underbrace{\boldsymbol{R}_{22}}_{4\times 4} \end{pmatrix}; \quad \underbrace{\boldsymbol{R}_2}_{8\times 8} := \begin{pmatrix} \underbrace{\boldsymbol{R}_{11}}_{4\times 4} & \underbrace{[\underbrace{\boldsymbol{R}_{12}}_{4\times 2} \,|\, \underbrace{\boldsymbol{0}}_{4\times 2}]}_{4\times 4} \\ \underbrace{\boldsymbol{0}}_{4\times 4} & \underbrace{\boldsymbol{I}}_{4\times 4} \end{pmatrix};$$

$$\boldsymbol{P}_2 = \boldsymbol{R}_1 \cdot \boldsymbol{R}_2 = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix};$$

$$\boldsymbol{P}_2^{-1}\boldsymbol{w}_t^* = \begin{bmatrix} 1 & 0 & -2 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -2 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{1,t-2} \\ y_{2,t-2} \\ y_{1,t-3} \\ y_{2,t-3} \\ y_{1,t-4} \\ y_{2,t-4} \end{bmatrix} = \underbrace{\begin{bmatrix} \Delta^2 y_{1,t-1} \\ \Delta^2 y_{2,t-1} \\ \Delta^2 y_{1,t-2} \\ \Delta^2 y_{2,t-2} \\ \Delta y_{1,t-3} \\ \Delta y_{2,t-3} \\ y_{1,t-4} \\ y_{2,t-4} \end{bmatrix}}_{\boldsymbol{w}_{t,2}}.$$

Now, let us define the following $N_{GC}p \times N_{GC}(p+d)$ matrix

$$\boldsymbol{M} := \left(\boldsymbol{I}_{N_{GC}p \times N_{GC}p} \quad \boldsymbol{0}_{N_{GC}p \times N_{GC}d}\right),$$

such that $\boldsymbol{\beta}_{\underline{GC}} = \boldsymbol{M}\boldsymbol{\beta}^*_{\underline{GC}} = \left(\boldsymbol{I}_{N_{GC}p \times N_{GC}p} \quad \boldsymbol{0}_{N_{GC}p \times N_{GC}d}\right) \begin{pmatrix} \underbrace{\boldsymbol{\beta}_{\underline{GC}}^{(1:p)'}}_{N_{GC}p \times 1} \\ \underbrace{\boldsymbol{\beta}_{\underline{GC}}^{(p+1:p+d)'}}_{N_{GC}d \times 1} \end{pmatrix}.$

Note that for $d = 1$

$$\boldsymbol{M}\boldsymbol{P}_1 = \boldsymbol{M}\boldsymbol{R}_1 = \tilde{\boldsymbol{R}}_{11} := \boldsymbol{R}_{11},$$

and for $d = 2$

$$\boldsymbol{M}\boldsymbol{P}_{\boldsymbol{2}} = \boldsymbol{M}\boldsymbol{R}_1\boldsymbol{R}_2 = \boldsymbol{M}\boldsymbol{R}_1^2.$$

In more detail, for $d = 1$ we get the reduced-upper-left matrix of $\boldsymbol{R}_1$:

$$\boldsymbol{M}\boldsymbol{P}_1 = \underbrace{\tilde{\boldsymbol{R}}_{11}}_{N_{GC}p \times N_{GC}p},$$

while for $d = 2$

$$\boldsymbol{M}\boldsymbol{P}_{\boldsymbol{2}} = \left(\boldsymbol{I}_{N_{GC}p \times N_{GC}p} \quad \boldsymbol{0}_{N_{GC}p \times N_{GC}2}\right) \begin{pmatrix} \underbrace{\boldsymbol{R}_{11}}_{N_{GC}p \times N_{GC}p} & \underbrace{\boldsymbol{R}_{12}}_{N_{GC}p \times N_{GC}2} \\ \underbrace{\boldsymbol{0}}_{N_{GC}2 \times N_{GC}p} & \underbrace{\boldsymbol{R}_{22}}_{N_{GC}2 \times N_{GC}2} \end{pmatrix} \times$$

$$
\begin{aligned}
&\times \begin{pmatrix} \underbrace{\boldsymbol{R}_{11}}_{N_{GC}p\times N_{GC}p} & \underbrace{[\ \underbrace{\boldsymbol{R}_{12}}_{N_{GC}p\times N_{GC}}\ \mid\ \underbrace{\boldsymbol{0}}_{N_{GC}p\times 2}\ ]}_{N_{GC}p\times N_{GC}2} \\ \underbrace{\boldsymbol{0}}_{N_{GC}2\times N_{GC}p} & \underbrace{\boldsymbol{I}}_{N_{GC}2\times N_{GC}2} \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{I}_{N_{GC}p\times N_{GC}p} & \boldsymbol{0}_{N_{GC}p\times N_{GC}2} \\ \boldsymbol{0}_{N_{GC}2\times N_{GC}p} & \boldsymbol{0}_{N_{GC}2\times N_{GC}2} \end{pmatrix} \begin{pmatrix} \underbrace{\boldsymbol{R}_{11}\boldsymbol{R}_{11}}_{N_{GC}p\times N_{GC}p} & \underbrace{\boldsymbol{R}_{11}[\boldsymbol{R}_{12}|\boldsymbol{0}]+\boldsymbol{R}_{12}}_{N_{GC}p\times N_{GC}2} \\ \underbrace{\boldsymbol{0}}_{N_{GC}2\times N_{GC}p} & \underbrace{\boldsymbol{R}_{22}}_{N_{GC}2\times N_{GC}2} \end{pmatrix} = \tilde{\boldsymbol{R}}_{11}^2 \\
&\equiv \boldsymbol{M}\boldsymbol{R}_1^2 = \begin{pmatrix} \boldsymbol{I}_{N_{GC}p\times N_{GC}p} & \boldsymbol{0}_{N_{GC}p\times N_{GC}2} \end{pmatrix} \begin{pmatrix} \underbrace{\boldsymbol{R}_{11}\boldsymbol{R}_{11}}_{N_{GC}p\times N_{GC}p} & \underbrace{\boldsymbol{R}_{11}\boldsymbol{R}_{12}+\boldsymbol{R}_{12}\boldsymbol{R}_{22}}_{N_{GC}p\times N_{GC}2} \\ \underbrace{\boldsymbol{0}}_{N_{GC}2\times N_{GC}p} & \underbrace{\boldsymbol{R}_{22}\boldsymbol{R}_{22}}_{N_{GC}2\times N_{GC}2} \end{pmatrix} \\
&= \tilde{\boldsymbol{R}}_{11}^2.
\end{aligned}
$$

Then, the following chain of equalities is verified:

$$
\boldsymbol{\phi}_{\underline{GC}} := \boldsymbol{M}\boldsymbol{\phi}^* = \boldsymbol{M}\boldsymbol{P}_d\boldsymbol{\beta}^*_{\underline{GC}} = \boldsymbol{M}\boldsymbol{R}_1^d\boldsymbol{\beta}_{\underline{GC}} = \tilde{\boldsymbol{R}}_{11}^d\boldsymbol{\beta}_{\underline{GC}}. \qquad (3.10)
$$

As $\boldsymbol{R}_{11}$ is invertible, it follows that any hypothesis formulated on $\boldsymbol{\beta}_{\underline{GC}}$ may equivalently be formulated in terms of $\boldsymbol{\phi}_{\underline{GC}}$ and vice-versa. Hence, by defining the function $f_d(\theta) := ((\tilde{\boldsymbol{R}}_{11}^d)^{-1}\theta)$ we just showed that testing the null hypothesis in (3.3) is equivalent of testing the null

$$
H_0: \quad f_d(\boldsymbol{\phi}_{\underline{GC}})\boldsymbol{S}_{N_J} = 0 \qquad \text{against} \qquad H_1: \quad f_d(\boldsymbol{\phi}_{\underline{GC}})\boldsymbol{S}_{N_J} \neq 0, \qquad (3.11)
$$

where $\boldsymbol{S}_{N_J}$ is a matrix of zeroes and ones, of conformable dimensions as $\boldsymbol{\phi}_{\underline{GC}}$ and which extracts only those coefficients corresponding to the variables in J. Therefore, e.g. the Lagrange Multiplier (LM) statistic for testing (3.11) gives the same numerical value as an LM test for testing (3.5). To show this it is sufficient to prove the numerical equivalence of the residual sum of squares (SSR_1) of the augmented regression (3.9) expressed in terms of $\hat{\boldsymbol{\phi}}_{\underline{GC}} = \boldsymbol{M}\hat{\boldsymbol{\phi}}^*$ with the residual sum of squares of regression (3.4), (SSR_2). First, we show that the equivalence chain in

(3.10) which is expressed in terms of the population parameter, it holds for the estimated version as well.

$$
\begin{aligned}
\boldsymbol{M}\hat{\boldsymbol{\phi}} &= \boldsymbol{M}\left(\boldsymbol{W}_d^{*'}\boldsymbol{G}_1\boldsymbol{W}_d^{*}\right)^{-1}\boldsymbol{W}_d^{*'}\boldsymbol{G}_1\boldsymbol{y}_I = \\
&= \boldsymbol{M}\boldsymbol{P}_d\left(\boldsymbol{X}_{\underline{GC}}^{\otimes *\prime}\boldsymbol{G}_1\boldsymbol{X}_{\underline{GC}}^{\otimes *}\right)^{-1}\boldsymbol{X}_{\underline{GC}}^{\otimes *\prime}\boldsymbol{G}_1\boldsymbol{y}_I \\
&= \boldsymbol{M}\boldsymbol{P}_d\hat{\boldsymbol{\beta}}^{*}_{\underline{GC}} = \boldsymbol{M}\boldsymbol{P}_d\begin{pmatrix}\hat{\boldsymbol{\beta}}^{(1:p)'}_{\underline{GC}} \\ \hat{\boldsymbol{\beta}}^{(p+1:p+d)'}_{\underline{GC}}\end{pmatrix} = \tilde{\boldsymbol{R}}^d_{11}\hat{\boldsymbol{\beta}}_{\underline{GC}},
\end{aligned}
\tag{3.12}
$$

for $\boldsymbol{G}_1 := \left[\boldsymbol{I} - \boldsymbol{X}^{\otimes}_{-GC}(\boldsymbol{X}^{\otimes\prime}_{-GC}\boldsymbol{X}^{\otimes}_{-GC})^{-1}\boldsymbol{X}^{\otimes\prime}_{-GC}\right]$.

Then, for the model in (3.9) by using results in (3.8), (3.10):

$$
\begin{aligned}
SSR_1 \equiv \hat{\boldsymbol{u}}'_I\hat{\boldsymbol{u}}_I &= \boldsymbol{y}'_I\boldsymbol{y}_I - 2\hat{\boldsymbol{\phi}}^{*'}\boldsymbol{W}_d^{*'}\boldsymbol{y}_I - 2\hat{\boldsymbol{\beta}}'_{-GC}\boldsymbol{X}^{\otimes'}_{-GC}\boldsymbol{y}_I + \\
&+ 2\hat{\boldsymbol{\phi}}^{*'}\boldsymbol{W}_d^{*'}\boldsymbol{X}^{\otimes}_{-GC}\hat{\boldsymbol{\beta}}_{-GC} + \\
&+ \hat{\boldsymbol{\phi}}^{*'}\boldsymbol{W}_d^{*'}\boldsymbol{W}_d^{*}\hat{\boldsymbol{\phi}}^{*} + \hat{\boldsymbol{\beta}}'_{-GC}\boldsymbol{X}^{\otimes'}_{-GC}\boldsymbol{X}^{\otimes}_{-GC}\hat{\boldsymbol{\beta}}_{-GC}, \\
&= \boldsymbol{y}'_I\boldsymbol{y}_I - 2\hat{\boldsymbol{\beta}}_{\underline{GC}}\boldsymbol{R}^d_{11}\boldsymbol{P}_d^{-1}\boldsymbol{X}^{\otimes}_{\underline{GC},d}\boldsymbol{y}_I - 2\hat{\boldsymbol{\beta}}'_{-GC}\boldsymbol{X}^{\otimes'}_{-GC}\boldsymbol{y}_I + \\
&+ \hat{\boldsymbol{\beta}}_{\underline{GC}}\boldsymbol{R}^d_{11}\boldsymbol{P}_d^{-1}\boldsymbol{X}^{\otimes}_{\underline{GC},d}\boldsymbol{X}_{-GC}\hat{\boldsymbol{\beta}}_{-GC} + \\
&+ \hat{\boldsymbol{\beta}}_{\underline{GC}}\boldsymbol{R}^d_{11}\boldsymbol{P}_d^{-1}\boldsymbol{X}^{\otimes}_{\underline{GC},d}\boldsymbol{P}_d^{-1}\boldsymbol{X}^{\otimes}_{\underline{GC},d}\hat{\boldsymbol{\beta}}_{\underline{GC}}\boldsymbol{R}^d_{11} + \hat{\boldsymbol{\beta}}'_{-GC}\boldsymbol{X}^{\otimes'}_{-GC}\boldsymbol{X}^{\otimes}_{-GC}\hat{\boldsymbol{\beta}}_{-GC}, \\
&= \boldsymbol{y}'_I\boldsymbol{y}_I - 2\hat{\boldsymbol{\beta}}^{*}_{\underline{GC}}\boldsymbol{P}_d\boldsymbol{P}_d^{-1}\boldsymbol{X}^{\otimes}_{\underline{GC},d}\boldsymbol{y}_I - 2\hat{\boldsymbol{\beta}}'_{-GC}\boldsymbol{X}^{\otimes'}_{-GC}\boldsymbol{y}_I + \\
&+ 2\hat{\boldsymbol{\beta}}^{*}_{\underline{GC}}\boldsymbol{P}_d\boldsymbol{P}_d^{-1}\boldsymbol{X}^{\otimes}_{\underline{GC},d}\boldsymbol{X}^{\otimes}_{-GC}\hat{\boldsymbol{\beta}}_{-GC} + \\
&+ \hat{\boldsymbol{\beta}}^{*}_{\underline{GC}}\boldsymbol{P}_d\boldsymbol{P}_d^{-1}\boldsymbol{X}^{\otimes}_{\underline{GC},d}\boldsymbol{P}_d^{-1}\boldsymbol{X}^{\otimes}_{\underline{GC},d}\hat{\boldsymbol{\beta}}^{*}_{\underline{GC}}\boldsymbol{P}_d + \hat{\boldsymbol{\beta}}^{\otimes'}_{-GC}\boldsymbol{X}^{\otimes'}_{-GC}\boldsymbol{X}^{\otimes}_{-GC}\hat{\boldsymbol{\beta}}^{\otimes}_{-GC}, \\
&= \boldsymbol{y}'_I\boldsymbol{y}_I - 2\hat{\boldsymbol{\beta}}^{*}_{\underline{GC}}\boldsymbol{X}^{\otimes}_{\underline{GC},d}\boldsymbol{y}_I - 2\hat{\boldsymbol{\beta}}^{\otimes'}_{-GC}\boldsymbol{X}^{\otimes'}_{-GC}\boldsymbol{y}_I + 2\hat{\boldsymbol{\beta}}^{*}_{\underline{GC}}\boldsymbol{X}^{\otimes}_{\underline{GC},d}\boldsymbol{X}^{\otimes}_{-GC}\hat{\boldsymbol{\beta}}'_{-GC} + \\
&+ \hat{\boldsymbol{\beta}}^{*'}_{\underline{GC}}\boldsymbol{X}^{\otimes'}_{\underline{GC},d}\boldsymbol{X}^{\otimes}_{\underline{GC},d}\hat{\boldsymbol{\beta}}^{*}_{\underline{GC}} + \hat{\boldsymbol{\beta}}'_{-GC}\boldsymbol{X}^{\otimes'}_{-GC}\boldsymbol{X}^{\otimes}_{-GC}\hat{\boldsymbol{\beta}}_{-GC} = \\
&= \boldsymbol{u}'_I\boldsymbol{u}_I \equiv SSR_2,
\end{aligned}
$$

where: $\hat{\boldsymbol{\beta}}_{-GC} = \left(\boldsymbol{X}^{\otimes'}_{-GC}\boldsymbol{G}_2\boldsymbol{X}^{\otimes}_{-GC}\right)^{-1}\boldsymbol{X}^{\otimes'}_{-GC}\boldsymbol{G}_2\boldsymbol{y}_I$ and
$\boldsymbol{G}_2 := \left[\boldsymbol{I} - \boldsymbol{W}_d^{*}(\boldsymbol{W}_d^{*'}\boldsymbol{W}_d^{*})^{-1}\boldsymbol{W}_d^{*'}\right]$and this shows the claim.

Remark 3.2. With the algebra presented in this section we are able to re-express the levels VAR in 3.1 by isolating the N_{GC} variables of interest for the causality testing from the potentially high-dimensional matrix of nuisance variables $\boldsymbol{X}^{\otimes}_{-GC}$. Furthermore, we can augment the lag-length of *only* the interest variables, incorporate these lag augmentations into the equation as in (3.4), such that we can re-state the interest variables in their d-differences (Δ^d) and this without implications on the null hypothesis. The inefficiency coming from the intentional overfitting of the VAR model by augmenting $p + d$ lags of *all* the variables included in the model, as original idea of Toda and Yamamoto (1995), gets greatly reduced even though the space of nuisance parameters is high-dimensional, as long as the interest is confined in testing sufficiently small portions of the variables in the system. Furthermore, as the variables of interest for the testing are expressed in their d-differences, this makes the asymptotic distribution of the testing procedures involving OLS estimators as e.g., Lagrange Multipliers and Wald test, to be standard χ^2. The rationale behind this is straightforward: even though we do not take the d differences of *all* the variables by means of augmenting d lags of them all, OLS χ^2 types tests of any linear hypothesis involving only the Δ^d-variables will converge at the usual parametric $\sqrt{T}$ rate, thus dominating the faster convergence of the non-stationary variables.

Remark 3.3. One important aspect of the current framework is that the lag-length p of the VAR is assumed to be larger than, or at most equal to, the suspected maximum order of integration d. This will be needed later in Section 3.3 to avoid spurious regression problems in the post-double selection steps but it is also of interest to observe here. In fact, one might be worried that having mixed orders of integration among the Granger caused and Granger causing variables could lead to over-differencing issues i.e., moving average unit roots being introduced by differencing stationary time series. This however does not happen here as the additional lags of the Granger causing and Granger caused variables are used "at convenience" i.e., if they are not needed because the variables are already stationary, at most they will marginally decrease the power of the test because of the slight over-specification of the lag-length but they will not affect the inference as the true coefficients of the extra lags are by construction always equal to zero.

Remark 3.4. The choice of $d = 2$ as augmentation in the above derivations and examples is done with purpose and is supported by simulations reported later in Section 3.5. It is well known that the distribution of the least squares estimator, when one or more roots of the characteristic polynomial are close to unity, becomes skewed yielding estimators which tend to underestimate the true autoregressive parameters (see e.g. Fuller, 2009). This skewness is also responsible to cause difficulties in constructing confidence intervals for the same parameters. Therefore, to avoid such near unit roots unwanted behaviors, augmenting $d = 2$ lags of the interest variables is always suggested. This does not cause overdifferencing issues as explained in Remark 3.3 and the simulations reported in Section 3.5 show satisfactory final sample behavior, also in terms of statistical power of the test.

3.3 Inference after selection by the lasso

We have shown in Section 3.2 that the augmentation of *only* the interest variables and the provided algebraic formulation allows to equivalently re-state the null hypothesis on $\boldsymbol{\beta}_{\underline{GC}}$ in terms of $\phi_{\underline{GC}}$. Appendix C contains the formal asymptotic justification for which the augmentation is needed to avoid non-standard limiting distributions of the test statistics in finite dimensions, when $K < T$. This is connected to the present context of a high-dimensional VAR model as in the proof of the main Theorem 3.1 in Section 3.4 we show with high probability that the set of retained variables within the double selection is close to the true, fixed dimensional support. We are now going to employ the post-double selection (PDS) LM test developed in Chapter 2 and adapt its theory and algorithm to the unit root non-stationary framework.

Consider again model (3.2). Let $\boldsymbol{x}_{GC,j}$, $j = 1, \dots, N_X$, where $N_X = pN_J$, denote the j-th column of $\boldsymbol{X}_{GC}$. Also, call $\boldsymbol{X}_{-GCj}$ the matrix $\boldsymbol{X}$

from which only the column corresponding to the j-th lag of $\boldsymbol{x}_J$ has been removed. Then, consider the "Frisch-Waugh" partial regressions of the variables of interest $\boldsymbol{y}_I$, $\boldsymbol{x}_{GC,j}$ on all other variables:

$$\boldsymbol{y}_I = \boldsymbol{X}^{\otimes}\boldsymbol{\eta}^{(0)} + \boldsymbol{e}^{(0)}, \tag{3.13}$$

$$\boldsymbol{x}_{GC,j} = \boldsymbol{X}_{-GCj}\boldsymbol{\eta}^{(j)} + \boldsymbol{e}^{(j)}, \qquad j = 1, \dots, N_X, \tag{3.14}$$

where $\boldsymbol{\eta}^{(j)}$, $j = 0, \dots, N_X$ are the best linear prediction coefficients for the prediction of respectively $\boldsymbol{y}_I$ and $\boldsymbol{x}_{GC,j}$ on $\boldsymbol{X}^{\otimes}$ and $\boldsymbol{X}^{(j)}_{-GCj}$. Consider the following scaling matrix

$$\boldsymbol{D}_T := \begin{pmatrix} \sqrt{T}I_A & 0 & 0 \\ 0 & TI_B & 0 \\ 0 & 0 & T^2 I_C \end{pmatrix},$$

and assume without loss of generality that the variables in $\boldsymbol{X}^{\otimes}$ and $\boldsymbol{X}_{-GCj}$ are organized by order of integration in increasing fashion, where A is the column dimension of the $I(0)$ variables, B is the same for $I(1)$ variables and C for $I(2)$ variables. Furthermore, define the limiting scaled expectation as $\bar{\mathbb{E}}(\cdot) := \lim_{T\to\infty} \boldsymbol{D}_T \mathbb{E}(\cdot)$. Then, for $j = 1, \dots, N_X$ the best linear predictions respectively for (3.13) and (3.14) are

$$\boldsymbol{\eta}^{(0)} = \arg\min_{\boldsymbol{\eta}} \bar{\mathbb{E}} \left\| \boldsymbol{y}_{I,t} - \boldsymbol{X}^{\otimes\prime}_{t-1}\boldsymbol{\eta} \right\|_2^2 = \left(\bar{\mathbb{E}} \boldsymbol{X}^{\otimes}_{t-1} \boldsymbol{X}^{\otimes\prime}_{t-1} \right)^{-1} \bar{\mathbb{E}} \boldsymbol{X}^{\otimes}_{t-1} \boldsymbol{y}_{I,t}, \tag{3.15}$$

$$\begin{aligned} \boldsymbol{\eta}^{(j)} &= \arg\min_{\boldsymbol{\eta}} \bar{\mathbb{E}} \left\| \boldsymbol{x}_{GC,j,t} - \boldsymbol{X}'_{-GCj,t-1}\boldsymbol{\eta} \right\|_2^2 = \\ &= \left(\bar{\mathbb{E}} \boldsymbol{X}_{-GCj,t-1} \boldsymbol{X}'_{-GCj,t-1} \right)^{-1} \bar{\mathbb{E}} \boldsymbol{X}_{-GCj,t-1} \boldsymbol{x}_{GC,j,t}, \end{aligned} \tag{3.16}$$

where $\boldsymbol{X}^{\otimes}_{t-1} = \boldsymbol{I}_{N_I} \otimes \boldsymbol{x}_{t-1}$. As $\boldsymbol{\eta}^{(0)}$ and $\boldsymbol{\eta}^{(j)}$ respectively obey the first order conditions: $\bar{\mathbb{E}}(\boldsymbol{y}_{I,t} - \boldsymbol{X}^{\otimes\prime}_{t-1}\boldsymbol{\eta})\boldsymbol{X}^{\otimes}_{t-1} = 0$,
$\bar{\mathbb{E}}(\boldsymbol{x}_{GC,j,t} - \boldsymbol{X}'_{-GCj,t-1}\boldsymbol{\eta})\boldsymbol{X}_{-GCj,t-1} = 0$, it follows that the errors $\boldsymbol{e}^{(0)}, \dots, \boldsymbol{e}^{(N_X)}$ are orthogonal to our variables of interest $\boldsymbol{y}_I$ and $\boldsymbol{x}_{GC,j}$.

Therefore, by partialling out the effects of all other variables guarantees a valid test of Granger causality. However, the sample versions of (3.13) and (3.14) are high-dimensional and cannot be directly estimated by least squares as $\boldsymbol{X}^{\otimes\prime}\boldsymbol{X}^{\otimes}$ and $\boldsymbol{X}'_{-GCj}\boldsymbol{X}_{-GCj}$ are not full rank. One needs to first select the relevant variables from regularized estimation of (3.13) and (3.14) and collect all these for the final estimation of $\boldsymbol{y}_I$ on $\boldsymbol{x}_{GC,j}$ plus only those relevant variables. By using the lasso (see Tibshirani, 1996) we can simultaneously perform variable selection and estimation of the parameters by solving respectively for (3.15) and (3.16) the following in-sample minimization problems

$$\begin{aligned}
\hat{\boldsymbol{\eta}}^{(0)} &= \arg\min_{\boldsymbol{\eta}}\left(T^{-1}||\boldsymbol{y}_I - \boldsymbol{X}^{\otimes}\boldsymbol{\eta}^{(0)}||_2^2 + \lambda||\boldsymbol{\eta}^{(0)}||_1\right), \\
\hat{\boldsymbol{\eta}}^{(j)} &= \arg\min_{\boldsymbol{\eta}}\left(T^{-1}||\boldsymbol{x}_{GC,j} - \boldsymbol{X}_{-GCj}\boldsymbol{\eta}^{(j)}||_2^2 + \lambda||\boldsymbol{\eta}^{(j)}||_1\right),
\end{aligned} \tag{3.17}$$

for $j = 1, \ldots, N_X$, where λ is a non-negative tuning parameter determining the strength of the penalty.

Remark 3.5. Kock (2016) showed that the (adaptive) lasso is oracle efficient in stationary as well as non-stationary autoregressions. One however needs to choose the tuning parameter λ appropriately as it needs to shrink the estimates of the truly-zero coefficients in $\boldsymbol{\eta}^{(0)}$, $\boldsymbol{\eta}^{(j)}$ to zero while at the same time it cannot grow too fast to avoid also the true non-zero parameters to be shrunk to zero. Here we follow the framework of Chapter 2 and use the Bayesian information criterion (BIC) in selecting the tuning parameter coupled with a penalty lower bound ensuring a maximum of selected variables per estimated equation (see Remark 3.8 for details). Minimizing an information criterion (IC) in order to determine an appropriate data-driven λ is one way to deal with dependent data (see Chapter 2 for an overview of other methods and their finite sample behaviors).

Let $\hat{S}(\lambda) = \{m \in \{1, \ldots, Kp\} : |\hat{\eta}_m(\lambda)| > 0\}$ denote the set of active variables in the lasso solution for a given λ. For a generic response vector $\boldsymbol{y}$ and predictor matrix $\boldsymbol{X}$, the value λ^{IC} is found as

$$\lambda^{IC} = \arg\min_{\lambda} \ln\left(\frac{1}{T}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\eta}}(\lambda)\|_2^2\right) + \frac{\ln T}{T}\left|\hat{S}(\lambda)\right|, \tag{3.18}$$

where $\left|\hat{S}(\lambda)\right|$ are the lasso degrees of freedom after the penalization procedure is applied i.e., the number of non-zero coefficients selected. The BIC-chosen λ within the adaptive lasso is well known to be able to identify the true model consistently as long as the model dimension is fixed (see e.g. Wang, Li, and Tsai 2007, Wang and Leng 2007). In fact, whenever the number of candidate models is fixed, BIC can consistently differentiate the true model from an arbitrary candidate model. However, if the model dimensions diverge, the number of candidate models increases at a too fast pace for the BIC to be able to distinguish the true model. Wang, Li, and Leng (2009) proposed a modified BIC and develop a set of probabilistic inequalities able to bound the overfitting coming from the diverging dimensions. The only difference with the standard BIC in (3.18) is that the penalty $\ln T/Tdf$ gets multiplied by a positive constant C_T. This constant is set to diverge to infinity but its rate can be arbitrarily slow, for instance Wang, Li, and Leng (2009) uses $C_T = \log(\log(K))$. Furthermore, a set of technical assumptions are needed in order to show that such modified BIC is consistent in model selection even with a diverging number of parameters. Specifically, (i) componentwise finite fourth-order moments for $\boldsymbol{X}$ are assumed, (ii) the minimal eigenvalue of the covariance matrix of $\boldsymbol{X}$ should be bounded away from zero (see also our Assumption 4,(g)), (iii) the divergence speed of the model dimension satisfy $\lim\sup(K/T^{\alpha}) < 1$ for $\alpha < 1$, and finally (iv) a limit requirement on the size of the non-zero coefficients is needed ($\sqrt{[T/C_T K \log(T)]}\liminf_{T\to\infty}(\min_{j\in S}|\boldsymbol{\eta}_j|) \to \infty$) as well as a constraint on the value of the diverging constant C_T ($C_T K \log(T)/T \to 0$). In our simulations in Section 3.5 we stick to the standard BIC, in fact simulations there show still a satisfactory performance of BIC without modifications. The theoretical argument of Wang, Li, and Leng (2009) is though appealing especially for ultra-high-dimensional settings. We compare it with standard BIC within the lag-length selection framework in Section 3.6.

Below in Algorithm 3 we state the main steps of our post-double selection, lag-augmented, Lagrange multiplier test and some remarks are included. We call the procedure PDS-LA-LM to stress the difference with the PDS-LM Algorithm 1 in Chapter 2. Here the post-selection step contains a lag-augmentation and this is restricted to the sole variables of interest for the testing.

Remark 3.6. In Algorithm 3, the choice among Step [4a] or [4b] does not affect the finite sample results of the test whenever the sample size T is large enough, similarly to Chapter 2. The small sample correction in [4b] (see Kiviet, 1986) has a wider practical applicability since [4a] suffers heavily for size distortion in small samples, therefore in Section 3.5 we always use [4b] for the Monte-Carlo simulations of the PDS-LA-LM test. The final sample results of using Step [4a] are reported for completeness in Table 3.9 in Appendix D.

Remark 3.7. As for Chapter 2, the feasible generalized least squares (FGLS) estimation in Step [3] of Algorithm 3 is needed when $N_I > 1$ to account for the correlation between equations of the VAR, and the fact the selected regressors are not the same in each equation. If $N_I = 1$, FGLS estimation boils down to the standard form of the LM statistic where one regresses $\hat{\boldsymbol{\xi}}$ by OLS onto the variables retained by the previous regularization steps plus the Granger causality variables, and retain the residuals $\hat{\boldsymbol{\nu}} = \hat{\boldsymbol{\xi}} - \boldsymbol{X}^{\otimes}_{\hat{S} \cup GC}\hat{\boldsymbol{\beta}}^{*}$, obtaining $R^2 = 1 - \hat{\boldsymbol{\nu}}'\hat{\boldsymbol{\nu}}/\hat{\boldsymbol{\xi}}'\hat{\boldsymbol{\xi}}$.

Algorithm 3 Post-double selection lag augmented Granger causality LM test (PDS-LA-LM)

[1] Estimate the initial partial regressions in (3.13) and (3.14) by an appropriate sparsity-inducing estimator such as the (adaptive) lasso, and let $\hat{\boldsymbol{\eta}}_0, \ldots, \hat{\boldsymbol{\eta}}_{N_X}$ denote the resulting estimators. Let $\hat{S}_0 = \{m : |\hat{\boldsymbol{\eta}}_{m,0}| > 0,\ m = 1, \ldots, N\}$ and $\hat{S}_j = \{m : |\hat{\boldsymbol{\eta}}_{m,j}| > 0,\ m = 1, \ldots, N_X\}$ for $j = 1, \ldots, p$, denote the selected variables in each regression.

[2] Let $\hat{S}_X = \bigcup_{j=1}^{N_X} \hat{S}_j$ denote all variables selected in the regressions for the columns of $\boldsymbol{X}_{GC}$, and let $\hat{S}_X^{\otimes}$ map $\hat{S}_X$ back to $\boldsymbol{X}^{\otimes}_{-GC}$ be such that $\boldsymbol{X}^{\otimes}_{\hat{S}_X^{\otimes}} = \boldsymbol{I}_{N_I} \otimes \boldsymbol{X}_{\hat{S}_X}$. Collect all variables kept by the lasso in Step [1] in $\hat{S}^{\otimes} = \hat{S}_0 \cup \hat{S}_X^{\otimes}$. Augment $\boldsymbol{X}^{\otimes}_{\hat{S}^{\otimes}}$ to $\widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}}$ including extra d lags of $\boldsymbol{y}_I$ and $\boldsymbol{x}_{GC}$. Obtain the residuals $\hat{\boldsymbol{\xi}} = \boldsymbol{y}_I - \widetilde{\boldsymbol{X}}^{\otimes}_{\hat{S}^{\otimes}} \hat{\boldsymbol{\beta}}^{\dagger}$ by OLS estimation. Let $\hat{\boldsymbol{\Xi}}_I$ denote the $T \times N_I$-matrix formed from $\hat{\boldsymbol{\xi}}$ and construct $\hat{\boldsymbol{\Sigma}}_{u,I} = \hat{\boldsymbol{\Xi}}_I' \hat{\boldsymbol{\Xi}}_I / T$ and $\hat{\boldsymbol{\Sigma}}^{\otimes}_{u,I} = \hat{\boldsymbol{\Sigma}}_{u,I} \otimes \boldsymbol{I}_T$.

[3] Let $\boldsymbol{y}^*_{N_I} = \left(\hat{\boldsymbol{\Sigma}}^{\otimes}_{u,I}\right)^{-1/2} \boldsymbol{y}_{N_I}$ and $\boldsymbol{X}^{*\otimes} = \left(\hat{\boldsymbol{\Sigma}}^{\otimes}_{u,I}\right)^{-1/2} \boldsymbol{X}^{\otimes}$. Obtain the residuals $\hat{\boldsymbol{\xi}}^* = \boldsymbol{y}^*_I - \boldsymbol{X}^{*\otimes}_{\hat{S}^{\otimes}} \hat{\boldsymbol{\eta}}^{\dagger}_{FGLS}$, and regress $\hat{\boldsymbol{\xi}}^*$ onto the variables retained by the previous regularization steps plus the Granger causality variables, retaining the residuals $\hat{\boldsymbol{\nu}}^* = \hat{\boldsymbol{\xi}}^* - \boldsymbol{X}^{*\otimes}_{\hat{S} \cup GC} \hat{\boldsymbol{\beta}}^*_{FGLS}$. Then obtain the statistic $LM = (\hat{\boldsymbol{\xi}}^{*\prime} \hat{\boldsymbol{\xi}}^* - \hat{\boldsymbol{\nu}}^{*\prime} \hat{\boldsymbol{\nu}}^*)$.

[4a] Reject H_0 if $LM > q_{\chi^2_{N_{GC}}}(1-\alpha)$, where $q_{\chi^2_{N_{GC}}}(1-\alpha)$ is the $1-\alpha$ quantile of the χ^2 distribution with $N_{GC} + d(N_I + N_J)$ degrees of freedom.

[4b] Reject H_0 if $\left(\frac{TN_I - \hat{s} - N_{GC} - d(N_I + N_J)}{N_{GC}}\right)\left(\frac{LM}{TN_{GC} - LM}\right) >$ $q_{F_{N_{GC}, N_I T - \hat{s} - N_{GC}} - d(N_I + N_J)}(1 - \alpha)$, where $\hat{s} = |\hat{S}^{\otimes}|$ and $q_{F_{N_{GC}, N_I T - \hat{s} - N_{GC}} - d(N_I + N_J)}(1 - \alpha)$ is the $1 - \alpha$ quantile of the F distribution with N_{GC} and $N_I T - \hat{s} - N_{GC} - d(N_I + N_J)$ degrees of freedom.

Remark 3.8. The algorithm designed in Chapter 2 employs a lower bound on the penalty to ensure that in each selection regression at most cT terms gets selected, for $0 < c < 1$. Similarly, we also employ a $c = 0.5$ upper-bound on the selected variables in Algorithm 3. This ensures that, equation-wise, the lasso does not select too many variables as this would render the union too large and hence infeasible for a post-least-squares estimator. Note that this is allowed as one of the main advantages of PDS is that it does not require consistent model selection but only prediction consistency (see Assumption 4,(e)). Mistakes are allowed to occur in the selection: variables might be incorrectly included and relevant variables might be missed, as long as the estimator remains sufficiently sparse and consistency is guaranteed. However, it remains a possibility that the lasso would select at every selection step an amount of variables correctly lower bounded, but substantially different for each step. The likelihood of this to happens obviously grows with the number of variables as well as the lags, although the latter we argue in Section 3.6 being reasonably assumed fixed (i.e., not growing with T or K) and small, in practice. In those limit cases where the selected variables are still larger than the sample size, post-OLS would remain infeasible both in an LM setting and similarly in an alternative, asymptotically equivalent, Wald test setting. The only work-around to these unfortunate cases is an ad-hoc increase for the tightness of the upper bound on the selected variables, namely from $c = 0.5$ to, say, $c = 0.5 - \tilde{\epsilon}$ for $\tilde{\epsilon} \in [0, 1]$, with the care of avoiding extreme tightness which would have implications for the testing.

Remark 3.9. Algorithm 3 is expressed in the general block-Granger causality notation. This shows that with this methodology one can test a block of variables being Granger causal for another variable or even for another block. However, differently from Chapter 2, here the testing must be confined to not-too-large portions of variables. The reason has to be found in the lag-augmentation framework of Section 3.2 needed to account for unit roots and cointegration. This essentially trades off power to reduce pre-test bias or bias occurring from taking the d-differences of the variables. The restricted augmentation developed in Section 3.2 essentially allows the lag-augmentation idea to work in high dimensions without sacrificing much power as it restricts the augmentation to only those variables of interest for the test. However, this lag augmentation still depends on the amount of variables involved in the hypothesis at stake. In Section 3.5 we show in simulations what is the difference in power loss if one augments all variables as opposed to the restricted augmentation of only the relevant ones for the testing.

3.4 Theoretical Results

This section presents the main theoretical contribution of this chapter. Throughout the section, in order to lighten the notation, we make the simplifying assumption that I contains a single element such that $N_I = 1$ i.e., we consider only one Granger caused variable, while we still allow for blocks among the Granger causing. Let us therefore introduce the notation $N_\varphi = N_J + N_I = N_J + 1$. The model then becomes:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} = \boldsymbol{X}_{\underline{GC}}\boldsymbol{\beta}_{\underline{GC}} + \boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC} + \boldsymbol{u}, \tag{3.19}$$

where $\boldsymbol{X}_{\underline{GC}}$ contains $N_\varphi p$ columns corresponding to the p lags of both Granger causing(s) and Granger caused variables. Similarly can be written the $N_X + 1$ selection steps in (3.13) and (3.14). Deriving theoretical results allowing for blocks in the Granger caused variables as well is straightforward from the theory presented here.

For the PDS-LA-LM to deliver uniformly valid inference, a set of assumptions are necessary. Especially, the assumptions of sparsity, restricted sparse eigenvalues, empirical process bound and consistency need to be adapted to the non-stationary framework. Hence, before showing that the post-double selection algorithm continues to deliver estimates free of omitted variable bias also in a non-stationary setting, we now state these assumptions and discuss them.

Assumption 4. Let δ_T and Δ_T denote sequences such that $\delta_T, \Delta_T \to 0$ as $T \to \infty$. Also, recall we assumed without loss of generality that elements in $\boldsymbol{X}$ are partitioned as $(\boldsymbol{X}_{A_{\hat{s}}}, \boldsymbol{X}_{B_{\hat{s}}}, \boldsymbol{X}_{C_{\hat{s}}})$ where columns of $\boldsymbol{X}_{A_{\hat{s}}}$ contains the $I(0)$ variables selected by the lasso ($\subseteq \boldsymbol{X}_{\hat{S}}$) as well as the Granger causing and dependent variables made stationary through the augmentation; $\boldsymbol{X}_{B_{\hat{s}}}$ contains the $I(1)$ and $\boldsymbol{X}_{C_{\hat{s}}}$ contains the $I(2)$ series. Conformably, recall the definition of the scaling matrix $\boldsymbol{D}_T = \text{diag}(\sqrt{T}\boldsymbol{I}_{A_{\hat{s}}}, T\boldsymbol{I}_{B_{\hat{s}}}, T^2\boldsymbol{I}_{C_{\hat{s}}})$. Then assume the following conditions are satisfied:

(a) **Martingale errors:** the error vector $\boldsymbol{e}$ in (3.13) and (3.14) is a K-dimensional martingale difference sequence with $\mathbb{E}(\boldsymbol{e}\boldsymbol{e}') = \boldsymbol{\Sigma}_e$ where:

  (i) For $\delta > 2$, $\mathbb{E}(||\boldsymbol{e}||_\infty^\delta) < C_\delta$, for some constant C_δ which depends on δ.

  (ii) There exist constants $c, C > 0$ such that: $c \leq \lambda_{\min}(\boldsymbol{\Sigma}_e) < \lambda_{\max}(\boldsymbol{\Sigma}_e) \leq C$

(b) **Limit Behavior:** given $\boldsymbol{e} = \boldsymbol{e}^{(0)}, \ldots, \boldsymbol{e}^{(N_X)}$ in (3.14), where $\boldsymbol{e}^{(j)}$ is defined in (a), then $\boldsymbol{D}_T^{-1}\boldsymbol{e}'\boldsymbol{u} \xrightarrow{d} N(0, \boldsymbol{\Omega})$ and $\boldsymbol{D}_T^{-1}\boldsymbol{e}'\boldsymbol{e}\boldsymbol{D}_T^{-1} \xrightarrow{p} \boldsymbol{\Sigma}_{GC|-GC}$, where

$$\begin{aligned}
\boldsymbol{\Omega} &= \lim_{T\to\infty} \boldsymbol{D}_T^{-1}\mathbb{E}\left(\boldsymbol{e}'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{e}\right), \\
\boldsymbol{\Sigma}_{GC|-GC} &= \lim_{T\to\infty} \boldsymbol{D}_T^{-1}\mathbb{E}\left(\boldsymbol{e}'\boldsymbol{e}\right)\boldsymbol{D}_T^{-1} = \\
&= \boldsymbol{\Sigma}_{GC,GC} - \boldsymbol{\Sigma}_{GC,-GC}\boldsymbol{\Sigma}_{-GC,-GC}^{-1}\boldsymbol{\Sigma}_{-GC,GC}.
\end{aligned}$$

(c) **Empirical Process:** with probability at least $1 - \Delta_T$

$$||\boldsymbol{D}_T^{-1}\boldsymbol{X}'\boldsymbol{u}||_\infty \leq 3\bar{\gamma}_T, \quad ||\boldsymbol{D}_T^{-1}\boldsymbol{X}'_{-GCj}\boldsymbol{e}||_\infty \leq 3\bar{\gamma}_T,$$

for $\bar{\gamma}_T$ being a bound applying to the three parts of $\boldsymbol{X}$ and $\boldsymbol{X}_{-GCj}$ and which depends on the sample size T as well as on several terms described in Theorem 3.4 in Appendix B.

(d) **Boundedness:** let $\boldsymbol{\beta}$ in (3.19) be in the interior of a compact parameter space $\mathbb{B} \subset \mathbb{R}^K$. Then, the (Granger causality) parameters of interest are bounded, that is, there exists a fixed constant $C > 0$ such that $\|\boldsymbol{\beta}_{GC}\|_1 \leq C$.[5]

(e) **Consistency:** with probability $1 - \Delta_T$, the lasso rate of convergence in prediction norm for $j = 0, \dots, p$ is given by

$$T^{-1}\left\|\boldsymbol{X}(\hat{\boldsymbol{\eta}}^{(j)} - \boldsymbol{\eta}^{(j)})\right\|_2^2 \leq \delta_T^2. \tag{3.20}$$

(f) **Sparsity:** let $S^{(j)} = \{\boldsymbol{\eta}_m \in \boldsymbol{\eta}^{(j)} : \boldsymbol{\eta}_m \neq 0\}$ be the sets of active variables in (3.13) and (3.14), and let $s = \left|\bigcup_{j=0}^{N_X} S^{(j)}\right|$ denote the cardinality of the set of all active variables (support), with $s = \{s_A, s_B, s_C\}$ where s_A contains the non-zero stationary variables, s_B the $I(1)$ and s_C the $I(2)$. The sparsity of the initial estimators is given by $\hat{s} = \{\hat{s}_A, \hat{s}_B, \hat{s}_C\} = |\hat{S}|$, where $\hat{S} = \bigcup_{j=0}^{p}\{\hat{\boldsymbol{\eta}}_m \in \hat{\boldsymbol{\eta}}^{(j)} : \hat{\boldsymbol{\eta}}_m \neq 0\}$. Then both the DGP and the estimator $\hat{\boldsymbol{\eta}}^{(j)}$ are sufficiently sparse; in particular, we have that with probability at least $1 - \Delta_T$, $\max(s, \hat{s}) \leq \bar{s}_T$ for some deterministic sequence $\bar{s}_T$.

(g) **Restricted Sparse Eigenvalues:** for any $\boldsymbol{\eta} \in \mathbb{R}^{(K-N_J)p}$ with $\|\boldsymbol{\eta}\|_0 \leq \bar{s}_T$, we have with probability at least $1 - \Delta_T$ that

$$\|\boldsymbol{\eta}\|_1 \leq \bar{s}_T \|\boldsymbol{X}\boldsymbol{\eta}\|_2 / \boldsymbol{\kappa}_{T,\min},$$

where $\boldsymbol{\kappa}_{T,\min} > 0$.

[5] $\boldsymbol{\beta}_{GC}$ is now the subvector of $\boldsymbol{\beta}_{\underline{GC}}$ corresponding to only the Granger causing variables.

(h) **Rate Conditions:** the interplay of the rates between the deterministic sequences for sparsity ($\bar{s}_T$), thickness of empirical process tails ($\bar{\gamma}_T$) and minimal eigenvalue ($\boldsymbol{\kappa}_{T,\min}$) yields:

$$T^3 \frac{\bar{s}_T \bar{\gamma}_T}{\boldsymbol{\kappa}_{T,\min}} \leq \delta_T. \tag{3.21}$$

Condition (a) is standard. In (i) we require at least two moments of the martingale difference sequence to exist while in (ii) we require minimum and maximum eigenvalues of the error covariance matrix to be bounded away from zero such that the matrix is non-singular. This restriction only rules out global dependence of the elements in $\boldsymbol{e}$ allowing for a wide range of contemporaneous dependencies.

Condition (b) assumes that a central limit theorem and weak law of large numbers hold. See e.g. Davidson (1994) for an overview of the various conditions under which these apply. Essentially the process should be sufficiently well-behaved in terms of moments and dependence allowed as we stated in (a). Note that we do not require iid-ness of the VAR error terms and we only need martingale difference errors.

Condition (c) bounds the empirical process with high probability. This uniform empirical process bound is obtained by a novel Gaussian approximation for martingale difference sequences applied to the three blocks of $\boldsymbol{X}'\boldsymbol{u}$ corresponding to the different integration orders of the time series in $\boldsymbol{X}$. Theorem 3.4 in Appendix B derives a Gaussian approximation for a general martingale difference sequence error term $\boldsymbol{\epsilon}$. This applies to both $\boldsymbol{u}$ and $\boldsymbol{e}$. The proof is presented in Appendix B after a sequence of preparatory lemmas.

Remark 3.10. In order to use the uniform approach in (c), typically an invariance principle is invoked such that every component of the vector $\boldsymbol{X}'\boldsymbol{u}^{(j)}$ is approximately Gaussian with negligible approximating error and hence standard sub-Gaussian tail bounds can be applied to show the claim. For instance, Condition 2 of Zhang, Robinson, et al. (2019), gives the following normal approximation result as $T \to \infty$ and for $0 < \tau < 1/2$, $\max_{1 \leq i \leq K} \max_{0 \leq \ell \leq 1} \mathbb{E}\left[\sum_{t=1}^{[T\ell]}(u_{i,t} - \sigma_{ii}\nu_{i,t})\right]^2 = \mathcal{O}(T^{2\tau})$ where $i = 1, \ldots, K$ indicates the elements of $\boldsymbol{u}$ i.e., mds assumed to be mean-zero[6], $\nu_{i,t}$ is an independent and standard normal sequence, $b_1 \leq \sigma_{ii}^2 \equiv \lim_{T\to\infty} \mathbb{V}ar\left(\sum_{t=1}^{T} u_{i,t}\right)/T \leq b_2$ $\forall i$ and b_1, b_2 positive constants. Note that the invariance principle of Zhang, Robinson, et al. (2019) is directly implied if the components of $\boldsymbol{u}$ are independent of each other and each component is an mds with a bit more than two finite moments. Given condition (a) and the fact that we can rewrite the non-stationary parts of $\boldsymbol{X}$ as vector moving averages (see equations (3.31), (3.32) in Appendix A) this would also hold for our case at least for the $I(0)$ and $I(1)$ parts. However, as the invariance principle result in Zhang, Robinson, et al., 2019 should hold separately for the three parts of $\boldsymbol{X}$, then the order of the approximating error would accumulate, thus aggravating the rates. Our result, as clear from the Theorem 3.5 in Appendix B is tighter, does not use any invariance principle and the probabilistic bounds we derive do not require explicit restrictions on the growth rate of K.

Condition (d) is standard and assumes compactedness of the parameter space of the vector $\boldsymbol{\beta}$ which in turn implies the boundedness.

Condition (e) is strictly related, and in fact follows, from Condition (f) and (g). Such inequality gives the rate of convergence of the lasso estimator in prediction norm. To satisfy the rates in (e), the upper bound can be shown through standard oracle inequality arguments (see e.g. Kock and Callot, 2015) to depend on the tuning parameter λ, the cardinality s of the active set, the restricted (sparse) eigenvalue as in

[6] This is without loss of generality as when the mean is non-zero the sequence of partial sums of u is not a mds but it is enough to center the partial sums by subtracting the mean.

(g) and some constants. We verify this in Appendix B.

Condition (f) requires sparsity of the DGP and the estimator. Sparsity of the first-stage estimator is needed in our framework as we perform OLS on the selected variables from the first-stage regressions. If the selected variables are not sparse enough, too many variables will be selected for OLS to be feasible. The assumption of exact sparsity in the DGP for the initial regressions can also be relaxed to approximate sparsity as in Belloni, Chernozhukov, and Hansen (2014b). Note that given the relevant d-augmentations discussed in Section 3.2, $\hat{s}_A \geq p$, $\hat{s}_{B,C} \geq d$.

Condition (g) is a key condition related to the appropriately scaled Gram matrices $\hat{\boldsymbol{\Sigma}}_i = \boldsymbol{X}_i'\boldsymbol{X}$, $i = (A_{\hat{s}}, B_{\hat{s}}, C_{\hat{s}})$. Whenever $K > T$, $\hat{\boldsymbol{\Sigma}}_i$ are degenerate i.e., $\min_{\boldsymbol{x} \in \mathbb{R}^K : \boldsymbol{x} \neq 0} \frac{\left(\boldsymbol{x}\hat{\boldsymbol{\Sigma}}_i\boldsymbol{x}\right)^{1/2}}{|\boldsymbol{x}|_2} \equiv \min_{\boldsymbol{x} \in \mathbb{R}^K : \boldsymbol{x} \neq 0} \frac{|\boldsymbol{X}_i\boldsymbol{x}|_2}{\sqrt{T}|\boldsymbol{x}|_2} = 0$, thus making OLS infeasible. Since directly imposing positive definiteness of $\hat{\boldsymbol{\Sigma}}_i$ i.e., $\min_{\boldsymbol{x} \in \mathbb{R}^K : \boldsymbol{x} \neq 0} \frac{|\boldsymbol{X}_i\boldsymbol{x}|_2}{\sqrt{T}|\boldsymbol{x}|_2} > 0$ would be a too strong assumption, Bickel, Ritov, et al. (2009) simply observed that for the lasso the minimum of the Rayleigh-Ritz quotient can instead be taken over a smaller set than the whole $\mathbb{R}^K$ i.e., any (non-zero) $T \times 1$ vector $\boldsymbol{x}$ such that $||\boldsymbol{x}_S||_1 \leq 3||\boldsymbol{x}_{S^c}||_1$ for S the true support and S^c its complement. The following quantity $\boldsymbol{\kappa}$ called the *restricted minimal eigenvalue* is therefore defined:

$$\boldsymbol{\kappa}_{\hat{\boldsymbol{\Sigma}}_i}(s, c) := \min_{\substack{S \subseteq \{1,\ldots,K\} \\ |S| \leq s}} \min_{\substack{\boldsymbol{x} \in \mathbb{R}^K \setminus \{0\} \\ ||\boldsymbol{x}_{S^c}||_1 \leq c||\boldsymbol{x}_S||_1}} \frac{\boldsymbol{x}'\hat{\boldsymbol{\Sigma}}_i\boldsymbol{x}}{\left\|\boldsymbol{x}_S\right\|^2}. \tag{3.22}$$

The cone condition allows to obtain rates of the model estimation error in ℓ_1-norm and hence those of the prediction loss using lasso. Any full-rank Gram matrix satisfies (3.22), therefore the population covariance matrix of the stationary variables in $\boldsymbol{X}$ i.e., $\boldsymbol{\Sigma}_{A_{\hat{s}}} = \mathbb{E}(\boldsymbol{X}_{A_{\hat{s}}}\boldsymbol{X}'_{A_{\hat{s}}})$, assumed being full-rank, automatically satisfies the condition. In Lemma

3.4 in Appendix B we show how a high probability bound on the maximal entrywise closeness between $\boldsymbol{\Sigma}_{A_{\hat{s}}}$ and the Gram matrix counterpart can be easily obtained. For the non-stationary case, since the covariance matrix is not a workable object, we rely on the results in Smeekes and Wijler (2021) who show that by allowing for a factor s to inflate the Gram matrix, it follows for $i = (B_{\hat{s}}, C_{\hat{s}})$, $\boldsymbol{\kappa}_{\hat{\boldsymbol{\Sigma}}_i}(s, c) > 0$ on a set with probability converging to one. Then one has that uniformly over all subsets S of cardinality at most s, with probability at least $1 - \Delta_T$ as $T, K \to \infty$, there exists a positive constant $\delta_T \neq 0$ such that

$$(I)\ \boldsymbol{\kappa}_{\hat{\boldsymbol{\Sigma}}_{A_{\hat{s}}}}(s,c) \geq \delta_T, \quad (II)\ \boldsymbol{\kappa}_{\hat{\boldsymbol{\Sigma}}_{B_{\hat{s}}}}(s,c) \geq \delta_T, \quad (III)\ \boldsymbol{\kappa}_{\hat{\boldsymbol{\Sigma}}_{C_{\hat{s}}}}(s,c) \geq \delta_T, \quad (3.23)$$

where: $\hat{\boldsymbol{\Sigma}}_{A_{\hat{s}}} = T^{-1}\boldsymbol{X}'_{A_{\hat{s}}}\boldsymbol{X}_{A_{\hat{s}}}$, $\hat{\boldsymbol{\Sigma}}_{B_{\hat{s}}} = T^{-2}s^2 \log^2 K(\boldsymbol{X}'_{B_{\hat{s}}}\boldsymbol{X}_{B_{\hat{s}}})$,
$\hat{\boldsymbol{\Sigma}}_{C_{\hat{s}}} = T^{-4}s^2 \log^2 K(\boldsymbol{X}'_{C_{\hat{s}}}\boldsymbol{X}_{C_{\hat{s}}})$ are the scaled Gram matrices where for the unit root cases they have been scaled up by a factor of $s^2 \log^2 K$ assumed to converge to zero as $s, K, T \to \infty$. It follows, $\boldsymbol{\kappa}_{\hat{\boldsymbol{\Sigma}}}(\boldsymbol{s}, c) > 0$ is assumed with probability $1 - \Delta_T$. $\boldsymbol{\kappa}_{T,\min}$ is therefore a positive constant satisfying (I-III) and which depends on the sample size T, the sparsity s and some constant c.

Finally, condition (h) is what links the sparsity $\bar{s}_T$, the tails thickness of the empirical process $\bar{\gamma}_T$ and the minimal eigenvalue $\boldsymbol{\kappa}_{T,\min}$. Since in Theorem 3.4 we proved a Gaussian approximation over the innovations $u_{i,t}$, it follows, as in Kock and Callot (2015), that $\bar{\gamma}_T$ could be taken of the order $\sqrt{\ln(K^2 p)}$ thus allowing for either fairly non-sparse models or almost exponentially large K with respect to T.

Remark 3.11. The current rate reported in (h) for the interplay between sparsity, tails thickness of the empirical process and minimal eigenvalue is suboptimal as it requires a factor T^3 to multiply the ratio in order for it to be bounded by an asymptotically vanishing sequence. This is an artifact of the current proof technique in Appendix C, where having elements of different orders in D_T leads to complications when taking norms, and means the norm of D_T and its inverse do not cancel out. We postulate that a different proof technique can prevent this issue and thus improve the rates. This is however outside the scope of this thesis.

Consider now the post selection lag-augmented equation as in Step [2] of Algorithm 3. We slightly deviate from Algorithm 3 as we directly include the p lags of the Granger causing variable(s) in the post-selection equation:

$$\boldsymbol{y} = \boldsymbol{X}^*_{\underline{GC}}\boldsymbol{\beta}^*_{\underline{GC}} + \boldsymbol{X}_{\hat{S}}\boldsymbol{\beta}_{\hat{S}} + \boldsymbol{u}, \tag{3.24}$$

where recalling $N_\varphi = N_J + 1$ then $\boldsymbol{X}^*_{\underline{GC}}$ is now the the $T \times N_\varphi(p+d)$ submatrix of $\boldsymbol{X}$ containing the original p lags of both the Granger causing and the Granger caused variables, as well as their additional augmented d lags, where the presence of the augmented elements is denoted with a $*$. Instead, $\boldsymbol{X}_{\hat{S}}$ denotes the the $T \times \hat{s}p$ submatrix of $\boldsymbol{X}$ corresponding to the p lags of the selected variables at Step [1] of Algorithm 3. The lags of the Granger caused contained in $\boldsymbol{X}^*_{\underline{GC}}$, i.e., the $p+d$ lags of $\boldsymbol{y}$, are needed from the theory developed in Section 3.2 for the definition of Granger causality. In what follows we will refer to $\boldsymbol{\beta}^*_{GC}$ as the subvector of $\boldsymbol{\beta}^*_{\underline{GC}}$ only containing the coefficients relative to the N_J variables.

Remark 3.12. In (3.24) we are assuming without loss of generality that immediately after the double selection in Step [1] of Algorithm 3, one would directly plug back the p lags of the Granger causing variable and use a Wald test on those coefficients, where the notation PDS denotes that the coefficients refer to the variables selected by the lasso at Step [1]. On the one hand, the choice of deriving the asymptotic normality for the PDS estimator $\hat{\boldsymbol{\beta}}^*_{GC}$ from a Wald test setting has the advantage of avoiding extra complications in the proof which would not add anything more insightful to the claim. On the other hand, the choice of stating Algorithm 3 in terms of the LM test in place of the Wald or the Likelihood Ratio (LR), has some practical advantages. It is well known that Wald and LM tests are asymptotically equivalent[7]

[7] Wald, LM and LR are asymptotically equivalent (see e.g. Engle, 1984, for a full treatment).

(we show this in Appendix A) and only slightly differ in finite samples (see e.g. Engle, 1984). In fact, they can even be written in equivalent form (See Appendix A, equation 3.35) with the only difference being the error covariance matrix which is calculated from the restricted least squares for the LM test, as opposed to the unrestricted least squares for the Wald test. It is also well established (see e.g. Savin, 1976) how under linear models, the Wald test will be numerically larger than, or at most equal to the LM one. In other words, whenever LM rejects the null, so does the Wald and whenever the Wald fails to reject so does the LM. As this is true under the null, the size of LM will necessarily be smaller or at most equal to that of Wald. Hence, in finite samples, the use of LM opens up to a more conservative testing procedure as opposed to a more liberal one in the case of the Wald. Even though this is relevant only for small samples and size corrections exists in the literature (see e.g. Rothenberg, 1982), still there are cases where the sample might be substantially small and smaller than the number of covariates. There, the use of LM could improve control of type I error. However, as observed later in our Monte Carlo study in Section 3.5, the d lags augmentation confined to only Granger causing and Granger caused has the important effect that the power of the testing procedure is not overly affected by the overspecification. However, power will be slightly affected even if one tests bivariate Granger causality. This is indeed the trade-off proposed by our method: giving up a little power in order to avoid biases from high-dimensional unit root and cointegration pre-testing. If the practitioner is more concerned with the power of the testing procedure, then employing a more liberal procedure as the Wald might result in slightly narrower confidence intervals. Note though that using the Wald test is not by any mean a way of enhancing the power of the test, this is in fact only a consequence of the more liberal nature in finite samples. Both in Chapter 2 and in some unreported finite sample exercises we do not find if minimal differences in the power of our Granger causality test, whether Wald or LM is used.

We are now going to state the main asymptotic result of this section, namely that the estimated post-double selection d-augmented least squares estimator for the coefficient vector $\boldsymbol{\beta}^*_{GC}$ is asymptotically Gaussian at the usual parametric rate.

Theorem 3.1. *Uniformly over a parameter space $\mathcal{S}$ for which Assumptions 4 (a)-(h) hold for all elements in $\mathcal{S}$,*

$$\sqrt{T}(\hat{\boldsymbol{\beta}}^*_{GC} - \boldsymbol{\beta}_{GC}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma}^{-1}_{GC|-GC}\boldsymbol{\Omega}\boldsymbol{\Sigma}^{-1}_{GC|-GC}), \qquad \text{as } T \to \infty.$$

The limiting distribution of the LM test is derived in the following Theorem 3.2.

Theorem 3.2. *Let $\boldsymbol{\beta}_{GC} = 0$. Then, uniformly over a parameter space $\mathcal{S}$ for which Assumption 4 (a)-(h) holds for all elements in $\mathcal{S}$ and for which $\boldsymbol{\Omega} = \sigma^2\boldsymbol{\Sigma}_{GC|-GC}$, where $\sigma^2 = \lim_{T\to\infty} \boldsymbol{D}_T^{-1}\mathbb{E}(\boldsymbol{u}'\boldsymbol{u})\boldsymbol{D}_T^{-1}$, we have that*

$$TR^2 \xrightarrow{d} \chi^2_p, \qquad \text{as } T \to \infty.$$

Heteroskedaticity-robust versions of the LM test could also be obtained at the price of some minor modifications of the test (see Wooldridge, 1987). We refer to Chapter 2, Algorithm 2 for a full treatment[8]. Proofs of Theorem 3.1 and 3.2 are reported in Appendix C.

3.5 Monte-Carlo Simulations

We now evaluate the finite-sample performance of our proposed PDS-LA-LM Granger causality test. Recall $\boldsymbol{y}_t = (y_{1,t}, \dots, y_{K,t})'$ and $\boldsymbol{u}_t =$

[8] Note that this is no different for the Wald test, for which the variance estimation has to be adjusted as well.

$(u_{1,t}, \ldots, u_{K,t})'$, then we consider the following Data Generating Processes (DGPs) in first differences inspired by Kock and Callot (2015):

$$\text{DGP1:} \qquad \Delta\boldsymbol{y}_t = \begin{bmatrix} 0.5 & 0 & \dots & 0 \\ 0 & 0.5 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0.5 \end{bmatrix} \Delta\boldsymbol{y}_{t-1} + \boldsymbol{u}_t,$$

$$\text{DGP2:} \qquad \Delta\boldsymbol{y}_t = \begin{bmatrix} (-1)^{|i-j|}a^{|i-j|+1} & \dots & (-1)^{|i-j|}a^{|i-j|+1} \\ (-1)^{|i-j|}a^{|i-j|+1} & \dots & (-1)^{|i-j|}a^{|i-j|+1} \\ \vdots & \ddots & \vdots \\ (-1)^{|i-j|}a^{|i-j|+1} & \dots & (-1)^{|i-j|}a^{|i-j|+1} \end{bmatrix} \Delta\boldsymbol{y}_{t-1} + \boldsymbol{u}_t,$$

with $a = 0.4$. The diagonal VAR(1) for DGP1 allows the sparsity assumption to be met. Instead, for DGP2 the coefficients decrease with exponential pace departing from the main diagonal and hence although the farthest coefficients are small, the sparsity assumption is not met. Note that we report simulations only for the bivariate Granger causality case where for simplicity we consider the first variable $(y_{1,t})$ in $\boldsymbol{y}_t$ being the Granger causing and the second $(y_{2,t})$ the Granger caused. Therefore, DGP1 automatically satisfies the null of no Granger causality from unit 2 to 1, however DGP2 does not. Therefore, we adapt DGP1 for the power analysis by setting the coefficient in position $(2,1)$ equal to 0.2. Conversely, we set the same coefficient equal to zero for DGP2 for the size analysis. We pick our time series of interest $y_{1,t}$ and $y_{2,t}$. For each DGP we are interested in the hypothesis that $y_{2,t}$ does not Granger cause $y_{1,t}$ i.e., for the VAR model as in 3.1 for $t = p+1, \ldots, T$

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \\ \vdots \\ y_{K,t} \end{bmatrix} = \sum_{j=1}^{p} \begin{bmatrix} a_{11}^{(j)} & a_{12}^{(j)} & a_{13}^{(j)} & \cdots & a_{1K}^{(j)} \\ a_{21}^{(j)} & a_{22}^{(j)} & a_{23}^{(j)} & \cdots & a_{2K}^{(j)} \\ a_{31}^{(j)} & a_{32}^{(j)} & a_{33}^{(j)} & \cdots & a_{3K}^{(j)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{K1}^{(j)} & a_{K2}^{(j)} & a_{K3}^{(j)} & \cdots & a_{KK}^{(j)} \end{bmatrix} \begin{bmatrix} y_{1,t-j} \\ y_{2,t-j} \\ y_{3,t-j} \\ \vdots \\ z_{K,t-j} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \\ u_{3,t} \\ \vdots \\ u_{K,t} \end{bmatrix},$$

then the tested null hypothesis is

$$H_0 : \ a_{21}^{(1)} = a_{21}^{(2)} = 0 \qquad \text{against} \qquad H_1 : \ a_{21}^{(j)} \neq 0, \ \text{for some } j = 1, 2.$$

Under the null and the alternative in turn, we integrate-out both DGP1 and DGP2 obtaining two VAR(2) in levels as

$$\text{DGP1:}\boldsymbol{y}_t = \begin{bmatrix} 1.5 & 0 & \dots & 0 \\ 0 & 1.5 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1.5 \end{bmatrix} \boldsymbol{y}_{t-1} + \begin{bmatrix} -0.5 & 0 & \dots & 0 \\ 0 & -0.5 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -0.5 \end{bmatrix} \boldsymbol{y}_{t-2} + \boldsymbol{u}_t,$$

$$\text{DGP2:}\boldsymbol{y}_t = \begin{bmatrix} 1+(-1)^{|i-j|}a^{|i-j|+1} & \dots & 1+(-1)^{|i-j|}a^{|i-j|+1} \\ 1+(-1)^{|i-j|}a^{|i-j|+1} & \dots & 1+(-1)^{|i-j|}a^{|i-j|+1} \\ \vdots & \ddots & \vdots \\ 1+(-1)^{|i-j|}a^{|i-j|+1} & \dots & 1+(-1)^{|i-j|}a^{|i-j|+1} \end{bmatrix} \boldsymbol{y}_{t-1} +$$

$$+ \begin{bmatrix} -(-1)^{|i-j|}a^{|i-j|+1} & \dots & -(-1)^{|i-j|}a^{|i-j|+1} \\ -(-1)^{|i-j|}a^{|i-j|+1} & \dots & -(-1)^{|i-j|}a^{|i-j|+1} \\ \vdots & \ddots & \vdots \\ -(-1)^{|i-j|}a^{|i-j|+1} & \dots & -(-1)^{|i-j|}a^{|i-j|+1} \end{bmatrix} \boldsymbol{y}_{t-2} + \boldsymbol{u}_t.$$

The lag-length is fixed to $p = 2$, namely the true lag-length for the non-stationary DGPs. For each non-stationary DGP we test with PDS-LA-LM the hypothesis that $y_{2,t}$ does not Granger cause $y_{1,t}$. Specifically, after the selection we employ a double ($d = 2$) augmentation of the dependent and the Granger causing variable as illustrated in Section 3.2. Following the recommendation in Chapter 2, we choose the BIC in selecting the tuning parameter λ for the lasso.

Table 3.1 reports the size and power of the PDS-LA-LM test out of 1000 replications. We use different combinations of time series length $T = (50, 100, 200, 500, 1000)$ and number of variables in the system $K = (10, 20, 50, 100)$ and a fixed lag-length $p = 2$. All the rejection frequencies are reported using a burn-in period of fifty observations. Simulations are also reported for different types of covariance matrices of the error terms. We employ a Toepliz-version for calculating the covariance matrix as $\Sigma_{i,j} = \rho^{|i-j|}$, where (i, j) refer to row i, column j of the matrix Σ_u. We cover two scenarios of correlation: $\rho = (0, 0.7)$. The first no-correlation is equivalent to set $\Sigma_{i,j} = I_{i,j}$, where I is the identity matrix.

Table 3.1: Simulation results for the PDS-LA-LM Granger causality test

DGP	Size/Power	ρ	K\T	50	100	200	500	1000
1	Size	0	10	8.3	8.2	5.3	3.9	3.9
			20	9.0	9.1	6.7	4.3	5.4
			50	8.8	7.2	8.4	5.0	4.5
			100	6.8	7.8	6.1	5.5	4.6
1	Power	0	10	19.8	40.6	78.1	99.8	100
			20	13.3	34.8	70.3	99.6	100
			50	14.3	31.1	64.1	98.7	100
			100	12.1	29.7	65.1	99.0	100
2	Size	0	10	7.6	7.7	4.9	5.6	5.1
			20	6.9	8.3	7.9	4.9	7.3
			50	7.1	6.4	5.7	7.0	6.3
			100	6.9	6.4	6.7	5.9	5.4
2	Power	0	10	15.5	27.4	50.2	94.8	99.9
			20	12.9	24.5	50.6	92.5	99.9
			50	11.2	24.1	44.7	90.3	99.9
			100	9.4	20.7	48.5	90.4	99.8
1	Size	0.7	10	10.2	7.1	5.3	5.5	4.9
			20	8.1	8.9	7.3	5.2	4.9
			50	8.1	7.2	9.2	7.4	5.4
			100	9.2	10.6	5.7	7.5	5.5
1	Power	0.7	10	15.6	21.9	39.8	85.1	99.4
			20	9.7	20.4	37.4	83.4	99.7
			50	11.1	19.0	33.5	79.8	98.5
			100	9.3	18.7	33.5	73.8	98.4
2	Size	0.7	10	9.5	7.6	5.2	6.4	7.8
			20	6.4	8.1	7.8	6.6	7.8
			50	7.4	8.2	8.1	7.6	7.1
			100	7.2	9.4	9.1	8.4	9.0
2	Power	0.7	10	10.9	19.7	34.1	78.1	98.3
			20	7.6	19.3	34.2	74.7	98.4
			50	9.6	21.7	32.8	71.2	97.9
			100	12.6	20.7	39.0	73.5	97.9

Notes: Size and Power for the different DGPs are reported for 1000 replications. $T = (50, 100, 200, 500)$ is the time series length, $K = (10, 20, 50, 100)$ the number of variables in the system, the lag-length is fixed to $p = 2$ and BIC is used to select the tuninig parameter for the lasso. ρ indicates the correlation employed to simulate the time series with the Toeplitz covariance matrix.

Our PDS-LA-LM test shows good performances in terms of size and (unadjusted) power for both DGPs considered. The setting of no correlation is handled remarkably well by both DGPs and only moderate size distortion is visible in large systems for small samples. Whenever high correlation of errors is present, sizes are still in the vicinity of 5% for DGP1 where the sparsity assumption is met. However, we notice how for DGP2, for which the sparsity assumption is not met, some residual size distortion remains visible even in large systems. However, the power of the test is always increasing with the sample size T for all the considered cases.

Remark 3.13. As mentioned in Remark 3.8, in order to obtain the results for the size and power when $T \leq Kp$ we need to impose a lower bound on the lasso penalty λ which guarantees to select at most cT variables in each relevant equation of the VAR, for some $0 < c < 1$. The bound should be set as strict as the system requires and often there is not a universal constant c that works in all settings, therefore this choice needs to be adaptive. For instance, if the lag-length is $p = 2$, this implies 3 selection steps (Step [1] of Algorithm 3). At the union of the selected variables, the d augmentated-lags of Granger caused and Granger causing variables are added (Step [2] of Algorithm 3) for a total of extra 4 more variables. The restrictiveness of the method used to tune the penalty in the lasso selection steps might in some cases not be sufficient to obtain $K < T$ before least squares. Only for these cases we tighten the bound using either $c = 0.33$ or $c = 0.25$.

Let us now elaborate on the main flow that the testing procedure designed in Toda and Yamamoto (1995) presents, namely the loss of power due to the inefficiency introduced by purposely overspecifying the VAR model with extra lags. We already mentioned in Section 3.2 how the fact that our proposed methodology involves only the augmentation of the variables of interest for testing causality, sensibly reduces the potential inefficiency. The original procedure suggested in Toda and Yamamoto (1995) augments d lags of all the regressors and was clearly envisioned for settings where the cross-sectional dimension was small. In high-dimensional systems, their procedure could be —if even feasible— potentially very inefficient. In fact, we could not possibly augment even one single extra lag of a full high-dimensional parameter vector, without making the system potentially intractable.

To show the difference in statistical power of our test, let us investigate the case of bivariate Granger causality for DGP2 with $T = (250, 500, 1000)$ and $K = (10, 20, 50, 100)$, namely the low-dimensional cases. We augment the system both as in Toda and Yamamoto (1995) with the third lag of all the variables and as we suggest (HMS) by augmenting two lags of only the Granger caused and the Granger causing variables. We compare the power of the test by also exploiting both a "pure" $I(1)$ case and a near-$I(2)$ case (see Remark 3.4). For the $I(1)$ case we modify the DGP diagonal elements $(\beta_{ii}^{(1)}, \beta_{ii}^{(2)})$ to $(\beta_{ii}^{(1)}, \beta_{ii}^{(2)}) = (1, 0)$ which returns the second highest eigenvalue of the companion matrix being equal to 0.533. Instead, for the near $I(2)$ case we use $(\beta_{ii}^{(1)}, \beta_{ii}^{(2)}) = (1.4, -0.4)$ which gives the second highest eigenvalue of the companion matrix being equal to 0.933.

Table 3.2: Power results

	$T = 250$		$T = 500$		$T = 1000$	
K	HMS	TY	HMS	TY	HMS	TY
10	57.5	54.1	90.1	88.8	99.6	99.2
20	54.4	47.9	86.4	82.8	99.9	99.7
50	50.6	31.2	85.2	77.3	99.7	99.3
100	52.5	NA	85.4	62.9	99.8	98.8

Notes: $(\beta_{ii}^{(1)}, \beta_{ii}^{(2)}) = (1, 0)$. HMS refers to our PDS-LA-LM test where we augment two lags of only dependent and Granger causing while TY refers to the LM test carried by augmenting all the regressors.

Table 3.3: Power results

	$T = 250$		$T = 500$		$T = 1000$	
K	HMS	TY	HMS	TY	HMS	TY
10	65.6	62.5	94.8	94.4	99.9	99.9
20	60.6	54.8	92.5	90.3	99.9	100
50	57.2	35.7	90.3	82.9	99.9	99.7
100	52.8	NA	90.4	68.3	99.8	99.1

Notes: $(\beta_{ii}^{(1)}, \beta_{ii}^{(2)}) = (1.4, -0.4)$. HMS refers to our PDS-LA-LM test where we augment two lags of only dependent and Granger causing while TY refers to the LM test carried by augmenting all the regressors.

Results in Table 3.2, 3.3 show that using our PDS-LA-LM test, the gain in statistical power is up to 30% for medium-to-small sample sizes if one augments only the dependent and the Granger causing variable, as opposed to augmenting all the regressors. NA's are reported when the after-selection lag augmentation made the system not feasible to be estimated with OLS. Results holds both in cases of "pure" I(1) and near-I(2) variables, showing that the $d = 2$ augmentation has practical advantages in the presence of near-unit roots and it does not dramatically decrease the power in those "pure" I(1) settings.

3.6 Choice of the lag-length p

Up until this point, we considered the lag-length p as given. In reality, this is of course not the case. In this section we elaborate on how we propose a reliable estimate of p. Note first that standard techniques for tuning the lag-length p as information criteria or sequential testing fail when applied directly to the full high-dimensional VAR. In fact, both techniques compare sequential estimates of VAR(p_i) models for a grid of lags, e.g., $i = \{1 : 10\}$. They then select the model either in a forward fashion by minimizing a chosen criteria like AIC, BIC, HQ or by backward testing the significance of the largest lag(s). This cannot simply be done in case of K being large and potentially larger than T. The addition of only one lag when K is large, quadratically inflates the parameter to estimates (K^2p) and can quickly lead to situations where $K > T$ and hence no OLS estimation is feasible. Shrinkage types of techniques could be considered to estimate p but they usually suffer from possible erratic behaviors due to high-correlations and depend on several ad-hoc choices as e.g., which estimation method to choose.

On another standpoint, there are theoretical reasons in favor of a small p, say from 1 to 2 in large VARs. As observed in Chapter 2, univariate ARMA models derived from a VAR(p) with for instance $p = 2$ lags and K $= 100$ series are already of maximal orders: ARMA$(Kp, (K - 1)p)$. Furthermore, also partial systems derived from the original VAR(p) will be VARMA of large order. Hence, given the usual estimated lag-length for macroeconomics application being $p = 4, 8$ for quarterly data with a small K, it is plausible to assume that the data generating process of the high-dimensional VAR has a small p.
Carrying forward this reasoning, we can calculate an empirical upper-bound on p by considering the $K \times K$ covariance matrix $\hat{\mathbf{\Omega}}$ obtained

using the residuals of a fitted diagonal VAR(p) as

$$\sum_{j=1}^{p}\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \\ \vdots \\ y_{K,t} \end{bmatrix} = \begin{bmatrix} a_{11}^{(j)} & 0 & 0 & \cdots & 0 \\ 0 & a_{22}^{(j)} & 0 & \cdots & 0 \\ 0 & 0 & a_{33}^{(j)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{KK}^{(j)} \end{bmatrix} \begin{bmatrix} y_{1,t-j} \\ y_{2,t-j} \\ y_{3,t-j} \\ \vdots \\ y_{K,t-j} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \\ u_{3,t} \\ \vdots \\ u_{K,t} \end{bmatrix}, \tag{3.25}$$

for $t = p+1, \dots, T$. Calling $\hat{\boldsymbol{u}}$ the $T \times K$ matrix of estimated residuals, then $\hat{\boldsymbol{\Omega}} := T^{-1}(\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}})$. Therefore, by using a grid of possible values for $p = \{1, \dots, 10\}$ we can select the model which minimise an information criterion (AIC,BIC) and hence get the upper bound lag-length p. Since the original lag-length should be small, this obtained value will not be far from the truth, if not exactly estimating the right p.
Let us consider the following information criteria[9]:

$$\begin{aligned} \text{AIC:} \quad & \log(det(\hat{\boldsymbol{\Omega}})) + \frac{2pK}{T}, \\ \text{BIC:} \quad & \log(det(\hat{\boldsymbol{\Omega}})) + \frac{\log(T)}{T}pK. \end{aligned}$$

Note that essentially this route allows to bypass the dimensionality issue in the equation-wise estimation, by considering the VAR coefficient matrix to be diagonal and hence estimating an AR(j) for each row of the VAR. This however solves half of the problem. In fact, when we build the covariance matrix $\hat{\boldsymbol{\Omega}}$, if the original set of variables K is larger than the sample size T available, $\hat{\boldsymbol{\Omega}}$ will be singular and hence we could not calculate any information criteria since these depends on the determinant of the covariance matrix which is equal to zero. In those cases where $K \geq T$, we can adopt an easy approximation for the determinant of $\hat{\boldsymbol{\Omega}}$, namely using the product of its diagonal elements; we further elaborate on this choice later in this section.

[9] Note that because we are estimating a diagonal VAR, AIC and BIC do not have K^2 but just K.

We build a simulation exercise to see how well this method performs and to compare the information criteria considered. First, we simulate using DGP1, DGP2 and different covariance specifications as in Section 3.5, a VAR(1) in first differences for various combinations of $K = (10, 20, 50, 100)$ and $T = (50, 100, 200, 500, 1000)$. We then integrate the series, thus obtaining a non-stationary VAR(2). Knowing the true value of $p = 2$, we apply the model selection procedure on a grid of 10 values for p. We report in Table 3.4, 3.5, 3.6 the percentage out of 100 replications of AIC and BIC selecting the right p. In the Appendix, Table 3.10, 3.11, 3.12 report the frequencies of the wrongly selected lag-lengths, namely the percentage out of 100 replications of AIC and BIC selecting models respectively with $p = 1$, $p = 3$ and $p > 3$ instead of $p = 2$.

Table 3.4: Selection of p, DGP1, $\rho = 0$

	$T = 50$		$T = 100$		$T = 200$		$T = 500$		$T = 1000$	
K	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
10	71	100	92	100	96	100	97	100	96	100
20	83	99	97	100	98	100	99	100	100	100
50	99	100	98	100	100	100	100	100	100	100
100	100	100	100	100	100	100	100	100	100	100

Notes: the values reported are percentage of correctly finding the true lag-length $p = 2$ out of 100 replications.

Table 3.5: Selection of p, DGP2, $\rho = 0$

	$T = 50$		$T = 100$		$T = 200$		$T = 500$		$T = 1000$	
K	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
10	58	96	65	100	34	98	0	82	0	22
20	68	73	74	100	12	100	0	57	0	0
50	85	100	80	100	13	100	0	45	0	0
100	93	100	36	100	36	100	0	63	0	0

Notes: the values reported are percentage of correctly finding the true lag-length $p = 2$ out of 100 replications.

Table 3.6: Selection of p, DGP2, $\rho = 0.7$

	$T = 50$		$T = 100$		$T = 200$		$T = 500$		$T = 1000$	
K	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
10	81	67	84	100	65	99	11	89	0	49
20	76	31	95	100	63	100	1	92	0	24
50	91	70	99	90	95	100	1	100	0	6
100	99	80	100	100	100	100	1	100	0	1

Notes: the values reported are percentage of correctly finding the true lag-length $p = 2$ out of 100 replications.

As expected, the empirical upper bound method works remarkably well with the diagonal, sparse DGP1. However, for DGP2 with an identity covariance matrix ($\rho = 0$), only BIC works satisfactorily and its good performance decreases with increasing T. Similarly, using DGP2 with a higher correlation structure ($\rho = 0.7$), the BIC is still preferable, however remarkably decreases its performance with higher T. Nevertheless, looking at the frequencies of the selected p in Table 3.10, 3.11, 3.12 we observe that when the system is large, BIC overestimates (mostly only one lag) the true lag-length. This is reassuring as it is much preferable to slightly overspecify the lag-length rather than underestimate it. We can take into account the overspecification (if not too large) without loosing too much efficiency in the testing. It turns out from Table 3.6 that the only occasion where AIC outperforms BIC is when the correlation is high and the system is small with still $T > K$. This is not surprising as the BIC in small finite samples suffers from being overly strict.

As earlier stated, we used the product of the diagonal element of $\hat{\mathbf{\Omega}}$ as an estimate for the determinant whenever $K \geq T$. This approach works quite well for both our considered DGPs. Therefore, we extend this particular estimation of the determinant to the whole simulation for DGP2 in Table 3.7 and 3.13.

Table 3.7: Selection of p, DGP2, $\rho = 0$

	$T = 50$		$T = 100$		$T = 200$		$T = 500$		$T = 1000$	
K	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
10	66	99	59	98	30	97	0	75	0	9
20	70	100	50	100	5	98	0	22	0	0
50	85	100	41	100	0	100	0	3	0	0
100	93	100	36	100	0	99	0	0	0	0

Notes: the values reported are percentage of correctly finding the true lag-length $p = 2$ out of 100 replications.

Again we observe a remarkable performance for BIC in small systems and especially whenever $K \geq T$. This good behavior of BIC however dissipates again and even earlier with increasing T. Nevertheless, observing the wrongly selected frequencies in Table 3.13 we notice, as before, a tendency of BIC to overestimate p and never to underestimate it. Furthermore, our simulation is a well-behaved scenario in terms of variable scales. In practice, when facing large datasets, the potentially huge scale differences, which cannot be mitigated by standardizations in the non-stationary context, might cause problems of near-singularity of $\hat{\mathbf{\Omega}}$. We find the determinant approximation to be able to circumvent this problem in large systems while the standard use of the determinant underestimates p. For these reasons we suggest the following version of BIC:

$$\begin{aligned} \text{BIC}^* : \quad & \log\left(\prod_{i=1}^{max(dim(\hat{\mathbf{\Omega}}))} (\hat{\mathbf{\Omega}}_{ii})\right) + \frac{\log(T)}{T} pK \\ & \equiv \operatorname{tr}\left(\log(\hat{\mathbf{\Omega}})\right) + \frac{\log(T)}{T} pK. \end{aligned} \tag{3.26}$$

for $\dim(\hat{\mathbf{\Omega}})$ being the row/column dimension of $\hat{\mathbf{\Omega}}$ and $\operatorname{tr}(\log(\hat{\mathbf{\Omega}}))$ being the trace of the log-transformed covariance matrix $\hat{\mathbf{\Omega}}$.

In light of Remark 3.5, it is sensible to analyze whether the BIC modification advocated by Wang, Li, and Leng (2009) in order to obtain BIC consistency even for diverging dimensions, outperforms (3.26) in finite samples. We then let $C_K = \ln(\ln(K))$ and define

$$\text{BIC}^{**}: \quad \operatorname{tr}\left(\log(\hat{\boldsymbol{\Omega}})\right) + \frac{\log(T)}{T} pKC_K. \tag{3.27}$$

In Table 3.8 we run again the same setting as in Table 3.7 where now BIC^{**} is used to select the lag-length

Table 3.8: Selection of p, DGP2, $\rho = 0$

	$T = 50$	$T = 100$	$T = 200$	$T = 500$	$T = 1000$
K	BIC**	BIC**	BIC**	BIC**	BIC**
10	100	97	91	48	1
20	100	100	100	38	0
50	100	100	100	57	0
100	100	100	100	72	0

Notes: the values reported are percentage of correctly finding the true lag-length $p = 2$ out of 100 replications.

BIC** performs better than BIC* for larger systems in terms of both T and K. Specifically, the main increased performance is observed for $T = 500$, otherwise the two are equivalent. For $T = 1000$ the lag-length is regularly overestimated for all K specifications. This is actually in line with Wang, Li, and Leng (2009): they in fact require the speed at which the dimension is allowed to diverge to be: $\limsup(K/T^{\kappa^*}) < 1$ for $\kappa^* < 1$. Under conditions reported in Remark 3.5, by Theorem 1,2 in Wang, Li, and Leng (2009) BIC* is then consistent even for diverging dimensions.

3.7 Empirical Application: Driving factors of Inflation

In this section we put into practice our developed framework in order to analyse the main driving factors of inflation in the US. We do this by means of creating Granger causality networks which can graphically represent the direction of the predicting connections from a set of macroeconomic variables to inflation. The interest in such an application is twofold. First, we show our procedure can be used in levels without needing to care about testing integration or cointegration for the series available in the dataset. Second, our procedure can be used as a preliminary step to identify instruments that can aid the inflation forecast. Forecasting inflation is clearly a very crucial task in rational economic decision-making. For instance, central banks rely on inflation forecasts to issue monetary policy as well as to fix inflation expectations to enhance policy efficacy. Inflation forecasts are also relevant for policymakers, businesses and households as contracts are normally issued in nominal terms. Here we use the FRED-MD dataset from McCracken and Ng (2016). Macroeconomics and econometrics literature have used this database intensively in the last five years, since it provides an excellent, up to date and structured source of "big data". We use the 12/2019 FRED-MD monthly release comprising a total of 128 variables, divided in eight macro topic-groups[10], sampled at monthly frequency from 01/01/1959 until 11/01/2019. We remove first the two initial rows corresponding to 01/01/1959 and 01/02/1959 in order to retain more series and facilitate later comparisons. We then remove those variables containing missing values. Thus the final dataset contains 107 macroeconomic variables for 729 datapoints. The lag-length estimated using the procedure outlined in Section 3.6 returns $p = 4$. We use "Consumer Price Index : All Items" (CPIAUCSL) as a proxy of inflation and we test all bivariate relations with the other 106 variables, each time conditioning on the remaining 105.

[10] "Output and income", "Labor market", "Housing", "Consumption, orders, and inventories", "Money and credit", "Interest and exchange rates", "Prices", "Stock market"

The FRED-MD database comes with a detailed appendix and Matlab/R routines which allow not only to clean the data from missings and outliers but also to take the appropriate transformations to render all the time series stationary. Although these tools facilitate the user and safeguard him from criticism over the variables handling, this is clearly rather an exception. Our developed tools (PDS-LA-LM) can be used directly on the raw data provided, without needing to take any difference directly and allowing the practitioner to pay no attention to the integration and cointegration properties of the time series at stake[11]. Also, by not taking differences we avoid the wiping off of the long memory features that some variables might exhibit. This is a particularly relevant aspect for macroeconomics and finance variables where the Box & Jenkins ARMA modeling approach of assuming the differentiated variables to be well behaved i.e., having fast decaying autocorrelations and trend-free, often fails (see e.g. Zivot and Wang, 2003).

To exploit this comparison, we run the same analysis on the stationary FRED-MD dataset where the time series have been transformed according to McCracken and Ng (2016). To test for Granger causality we use the same PDS algorithm designed in Chapter 2 for stationary time series (PDS-LM). The same algorithm to select the lag-length p as in Section 3.6 estimates this time $p = 1$. This is already interesting on its own right given that for the level case p was estimated equal to 4. By differentiating the variables the dynamic gets greatly reduced to only one lag, signifying already the consistent loss in memory of the series. We report the analysis both at significance level $\alpha = 0.05$ and $\alpha = 0.01$ for both cases.

[11] To run these analyses we used the authors R package “HDGCvar” available at `https://github.com/Marga8/HDGCvar`

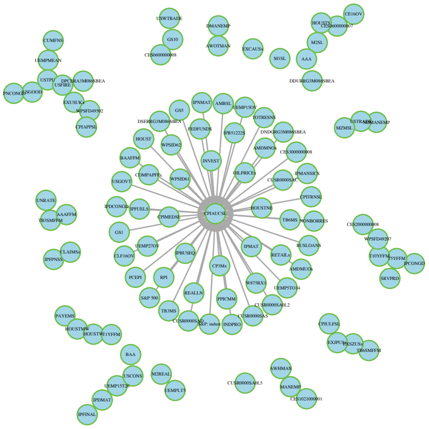


Figure 3.1: PDS-LA-LM, $\alpha = 0.05$

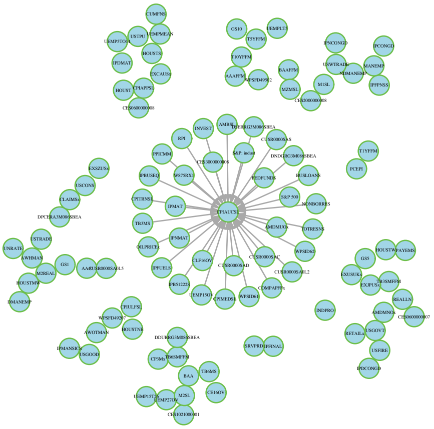


Figure 3.2: PDS-LA-LM, $\alpha = 0.01$

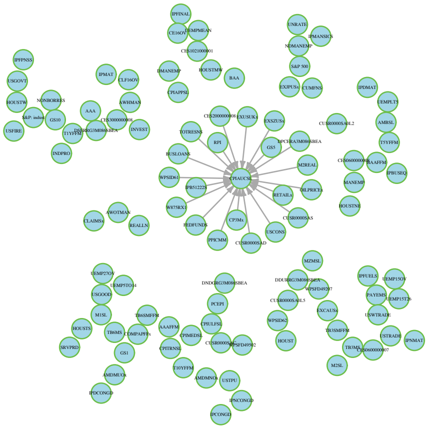


Figure 3.3: PDS-LM, $\alpha = 0.05$

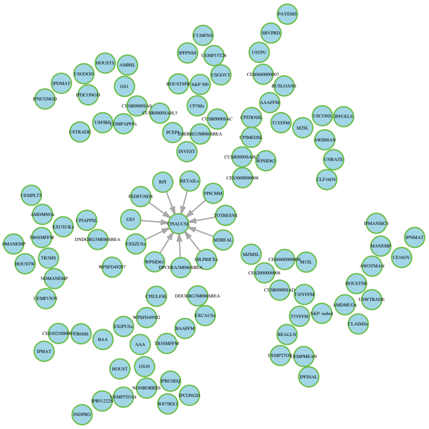


Figure 3.4: PDS-LM, $\alpha = 0.01$

Results in Figure 3.1, 3.2 refer to our PDS-LA-LM test performed on levels while those in Figure 3.3, 3.4 refer to the PDS-LM test in Chapter 2 on the stationary-transformed data according to McCracken and Ng (2016). First, we note how quantitatively the number of connections among macroeconomics variables and inflation is substantially higher when we consider the levels while several connections disappear when considering the differences. Specifically, at $\alpha = 0.05$ Figure 3.1 counts 50 connections. By tightening the significance to $\alpha = 0.01$, Figure 3.2 shows 33 connections. Correspondingly in the differences case in Figure 3.3, 3.4 are only 20 and 11. When considering the PDS-LA-LM test at $\alpha = 0.01$ the several connections shown belongs to the macro-variable groups: Money and credit: INVEST, AMBSL, BUSLOANS, TOTRESNS, NONBORRES, Output and Income: RPI, W875RX1, IPBUSEQ, IPMAT, IPNMAT, IPFUELS, IPB51222S, Prices: DSERRG3M086SBEA, CUSR0000SAS, WPSID62, OILPRICEx, CPITRNSL, CPIMEDSL, CUSR0000SAD, CUSR0000SAC, CUSR0000SA0L2, DNDGRG3M086SBEA, PPICMM, Interest and Exchange Rates: (FEDFUNDS, TB3MS, COMPAPFFx, Stock Market: S&P 500, S&P: indust, Labor Market: UEMP15OV, CES3000000008, CLF16OV, Consumption Orders and Inventories: AMDMUOx. All the found connections are sensible and economically justifiable: inflation and output are related as postulated in the Phillips curve (see e.g. connections with Real Personal Income and the industrial production indeces); inflation has also impact on financial institution, the markets (see e.g. the connections with S&P 500, S&P: indust) and has implications for investment policy. In accordance with the macroeconomic literature, oil price (OILPRICEx) is found as leading indicator of inflation, thus validating the pass-through theory (see e.g. Blanchard and Gali, 2007), which is a particularly important matter for monetary policy implementation. Of similar interest is the connection between both 3-Month Treasury Bill (TB3MS), Effective Federal Funds Rate (FEDFUNDS) with inflation. Treasury Bills are medium/short-term obligations issued by the U.S. Treasury Department, holding -in this specific case- a maturity of 3 months. Hence, they are short term, zero default risk investments, meant to increase in volume whenever the economy faces periods of uncertainty or stagnation. In these times investors prefer to buy more secure obligations rather than long term more productive ones, although they obtain less return from them. Treasuries are strictly connected with (expected) inflation; investors tend not to buy treasuries in periods when inflation rate is higher than the return of the obligation. Furthermore, treasury bill rates are often considered as proxy of expected inflation (see Fama and Schwert (1977)). When inflation increases, the price of treasuries usually decreases and our analysis also shows how these short term securities are able to be leading indicators of inflation.

If compared to Figure 3.4, the common connections are: RPI, FEDFUNDS, WPSID61, OILPRICEx, TOTRESNS, PPICMM. Five connections found in Figure 3.3, 3.4 are not common to either Figure 3.1 or 3.2, namely: EXSZUSx, DPCERA3M086SBEA, M2REAL, USCONS, CES2000000008. These connections might therefore be considered as spurious discoveries induced by the differencing.

3.8 Conclusion

We build an inferential procedure for Granger causality testing in high-dimensional non-stationary VAR models which avoids any integration or cointegration biased pre-test. To do so we adapt the Toda and Yamamoto (1995) idea of augmenting the lag-length of the system and we show that by reducing this augmentation to only the variables of interest for the testing we are able to sensibly diminish the efficiency loss coming from the model overspecification. To handle the high-dimensionality of the VAR we develop a post-double selection LM test which is based on penalized least square estimators. Using the lasso we are able to partial-out those variables having no influence in the tested relation while safeguarding from omitted variable bias using a double-selection mechanism. We present the algebra needed to prove that the augmentation of the interest variables has no effect on the null-hypothesis tested, thus letting the OLS estimator having standard asymptotic results, free of nuisance terms. Also, we extend the relevant assumptions needed for the post-double selection estimator to work in the context of potentially unit root non-stationarities. We derive the asymptotics of the post-selection augmented estimator, showing it attains standard asymptotic normality hence allowing for a valid LM test with standard χ^2 limiting distribution.

Our proposed test shows good finite sample properties over different DGPs, namely both sparse and non-sparse and with different covariance structures. We also give practical recommendations on both the optimal augmentation d and on how to estimate the lag-length p. We argue that $d = 2$ lags is the optimal augmentation in order to take into account possible I(2) as well as near I(2) variables that could compromise the right convergence of the test. We show that this has a minimal impact on the efficiency of the test since we are only required to augment the variables of interest for the causality test. In order to estimate the lag-length p, in the spirit of the observed fact that larger systems tend to be described by fewer lags, we propose to reduce the original VAR to a diagonal VAR. This takes care of the equation-wise potential dimensionality problem (whenever $K > T$) and reduces the system to a sequential feasible estimation of $\mathrm{AR}(j)$ equations for a grid of lags j. For each j we can then build the covariance matrix from the AR residuals, thereby obtaining a square covariance matrix to be used in the specification of information criteria which we minimize in order to select the correct lag-length. To avoid issues of singularity or near-singularity of the covariance matrix, we propose to estimate its log-determinant by means of the trace of the log-covariance matrix, thus re-defining the information criteria to use in the model selection.

Finally, we investigate how our test performs in practice by analysing the main driving factors of inflation in the US. Using the FRED-MD data set directly, without needing to apply their suggested transformations, we are able to derive causality networks connecting macroeconomic variables which lead inflation thus proving the usefulness of our method in finding valid instruments for inflation forecasting.

Appendix A Preparatory Lemmas

In this section we report some lemmas that will be used later in Appendix B, C to show the main results.

Consider the conditions developed in Johansen (1992) which guarantee the process to be I(1) or maximum I(2) and, in general, cointegrated. Consider rewriting the levels VAR equation (3.1) as

$$\boldsymbol{y}_t = \sum_{j=1}^{p} \boldsymbol{A}_j \boldsymbol{y}_{t-j} + \boldsymbol{u}_t, \tag{3.28}$$

where $\boldsymbol{A}_j$ are the coefficient matrices for each lag j.
First, explosive processes are ruled out:

Assumption 5.

$$\begin{aligned} &|\boldsymbol{A}(z)| = 0 \quad \text{implies} \quad |z| > 1 \quad \text{or} \quad z = 1, \\ &\text{where} \quad \boldsymbol{A}(z) = \boldsymbol{I}_T - \boldsymbol{A}_1 z - \cdots - \boldsymbol{A}_p z^p. \end{aligned}$$

Then we can re-express (3.28) in the VECM format

$$\Delta \boldsymbol{y}_t = \sum_{j=1}^{p-1} \tilde{\boldsymbol{A}}_j \Delta \boldsymbol{y}_{t-j} + \boldsymbol{\Pi}_p \boldsymbol{y}_{t-p} + \boldsymbol{u_t}, \tag{3.29}$$

where $\tilde{\boldsymbol{A}}_i = \sum_{h=1}^{i} \boldsymbol{A}_h - \boldsymbol{I}_T (i = 1, \dots, p-1)$ and $\boldsymbol{\Pi}_p = -\boldsymbol{A}(1)$.[12]

Assumption 6.

$$\boldsymbol{\Pi}_p = \boldsymbol{\mathcal{A}} \boldsymbol{B}',$$

for some $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{B}$, where $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{B}$ are $K \times r$ matrices of rank r.

[12] Out of simplicity we omitted the intercept as well as the linear trend term in the cointegrating relationship, the analysis goes trough in a similar way if they are included.

We also need:

Assumption 7.

$$\boldsymbol{\mathcal{A}}_{\perp}'\boldsymbol{\Pi}_1\boldsymbol{B}_{\perp},$$

is nonsingular, where $\boldsymbol{\Pi}_1 = -\tilde{\boldsymbol{A}}(1)$, with $\tilde{\boldsymbol{A}}(z) = \boldsymbol{I}_T - \tilde{\boldsymbol{A}}_1 z - \ldots - \tilde{\boldsymbol{A}}_{p-1}z^{p-1}$ and $\boldsymbol{\mathcal{A}}_{\perp}$ and $\boldsymbol{B}_{\perp}$ are $K \times (K-r)$ matrices of rank $K-r$ such that $\boldsymbol{\mathcal{A}}'\boldsymbol{\mathcal{A}}_{\perp} = \boldsymbol{B}'\boldsymbol{B}_{\perp} = 0$.

Under assumption 5-7, the process is $I(1)$ and cointegrated if $r > 0$.

We can further re-write (3.29) as VECM

$$\Delta^2\boldsymbol{y}_t = \sum_{j=1}^{p-2}\boldsymbol{A}_j^*\Delta^2\boldsymbol{y}_{t-j} + \boldsymbol{\Pi}_{p-1}\Delta\boldsymbol{y}_{t-p+1} + \boldsymbol{\Pi}_p\boldsymbol{y}_{t-p} + \boldsymbol{u}_t, \tag{3.30}$$

where $\boldsymbol{A}_i^* = \sum_{h=1}^{i}\boldsymbol{J}_h - \boldsymbol{I}_K (i = 1, \ldots, p-2)$, hence we have the following assumption.

Assumption 8.

$$\bar{\boldsymbol{\mathcal{A}}}_{\perp}'\boldsymbol{\Pi}_{p-1}\bar{\boldsymbol{B}}_{\perp} = \boldsymbol{F}\boldsymbol{G}',$$

for some $\boldsymbol{F}$, $\boldsymbol{G}$, where $\bar{\boldsymbol{\mathcal{A}}}_{\perp} = \boldsymbol{\mathcal{A}}_{\perp}\left(\boldsymbol{\mathcal{A}}_{\perp}'\boldsymbol{\mathcal{A}}_{\perp}\right)^{-1}$, $\bar{\boldsymbol{B}}_{\perp} = \boldsymbol{B}_{\perp}\left(\boldsymbol{B}_{\perp}'\boldsymbol{B}_{\perp}\right)^{-1}$ and $\boldsymbol{F}$ and $\boldsymbol{G}$ are $(K-r) \times s$ matrices of rank s $(0 < s < K-r)$.

Under assumption 5, 6, 8 and (2.8) of Johansen (1992) which prevents it from being $I(3)$, the process is $I(2)$ and it is cointegrated unless $r = s = 0$. Also, given assumption 5-8, the VECMs in (3.29), (3.30) can be rewritten according to the Granger representation theorem in their VMA format respectively as

$$\boldsymbol{y}_t = \boldsymbol{C}\sum_{i=1}^{t}\boldsymbol{u}_{\boldsymbol{i}} + \boldsymbol{C}(\boldsymbol{L})\boldsymbol{u}_t, \tag{3.31}$$

$$\boldsymbol{y}_t = \boldsymbol{C}_2 \sum_{i=1}^{t} \sum_{j=1}^{i} \boldsymbol{u}_j + \boldsymbol{C}_1 \sum_{i=1}^{t} \boldsymbol{u}_i + \boldsymbol{z}_t, \tag{3.32}$$

where $\boldsymbol{C} = \boldsymbol{B}_\perp(\boldsymbol{\mathcal{A}}'_\perp \boldsymbol{\Pi}_1 \boldsymbol{B}_\perp)^{-1}\boldsymbol{\mathcal{A}}'_\perp$, $\boldsymbol{C}_1, \boldsymbol{C}_2$ are functions of the model parameters and $\boldsymbol{z}_t$ is $I(0)$.

The following proposition is needed for the asymptotic analysis is Appendix B, C. The results contained are well known[13] and they have been derived in Phillips and Durlauf (1986), Park and Phillips (1988), Sims et al. (1990), Phillips and Solo (1992), therefore we state only those necessary to our analysis. As later in Appendix C, write the true model as

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{X}^*_{GC}\boldsymbol{\beta}^*_{GC} + \boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC} + \boldsymbol{u} \\ &= \boldsymbol{X}^*_{GC}\boldsymbol{P}_d\boldsymbol{P}_d^{-1}\boldsymbol{\beta}^*_{GC} + \boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC} + \boldsymbol{u} \\ &= \boldsymbol{W}^*_d\boldsymbol{\phi}^* + \boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC} + \boldsymbol{u},\end{aligned}$$

for $\boldsymbol{W}^*_d := \boldsymbol{X}^*_{GC}\boldsymbol{P}_d, \boldsymbol{\phi}^* := \boldsymbol{P}_d^{-1}\boldsymbol{\beta}^*_{GC}$.

Proposition 3.1. *Assume without loss of generality that $\boldsymbol{X}_{-GC}$ can be partitioned as $(\boldsymbol{X}_{-GC,A}, \boldsymbol{X}_{-GC,B}, \boldsymbol{X}_{-GC,C})$ and also that $(\boldsymbol{W}^*_{d,A}, \boldsymbol{W}^*_{d,B}, \boldsymbol{W}^*_{d,C}), (\boldsymbol{X}_{-GC,A}, \boldsymbol{X}_{-GC,B}, \boldsymbol{X}_{-GC,C})$ are given initial (joint) distribution such that $\left(\boldsymbol{W}^*_{d,A}, \Delta\boldsymbol{W}^*_{d,B}, \Delta^2\boldsymbol{W}^*_{d,C}\right)$, $\left(\boldsymbol{X}_{-GC,A}, \Delta\boldsymbol{X}_{-GC,B}, \Delta^2\boldsymbol{X}_{-GC,C}\right)$ are stationary. Then let,*[14]

$$\boldsymbol{\omega}_t := \left(\boldsymbol{u}_t, \boldsymbol{\tau}_{t,A}, \Delta\boldsymbol{\tau}_{t,B}, \Delta^2\boldsymbol{\tau}_{t,C}\right),$$

where $\boldsymbol{\tau}_{t,i} := \begin{pmatrix}\boldsymbol{W}^*_{d,i} \\ \boldsymbol{X}_{-GC,i}\end{pmatrix}$ for $i = (A, B, C)$. Then define for all t:

$$\boldsymbol{\Sigma} = \mathbb{E}\boldsymbol{\omega}_t'\boldsymbol{\omega}_t,$$

[13] For a full treatment of these results we refer to Hamilton (1994).

[14] Note that $\boldsymbol{W}^*_d$ and $\boldsymbol{X}_{-GC}$ are partitioned conformably but have different dimensions.

$$\mathbf{\Lambda} = \sum_{j=1}^{\infty} \mathbb{E}\boldsymbol{\omega_t'}\boldsymbol{\omega_{t+j}},$$

$$\mathbf{\Omega} = \mathbf{\Sigma} + \mathbf{\Lambda} + \mathbf{\Lambda}'.$$

Then, partition $\mathbf{\Sigma}, \mathbf{\Lambda}, \mathbf{\Omega}$ *conformably with* $\boldsymbol{\omega}_t$ *such that*

$$\mathbf{\Sigma} := \begin{pmatrix} \mathbf{\Sigma}_u & \mathbf{\Sigma}_{uA} & \mathbf{\Sigma}_{uB} & \mathbf{\Sigma}_{uC} \\ \mathbf{\Sigma}_{Au} & \mathbf{\Sigma}_{A} & \mathbf{\Sigma}_{AB} & \mathbf{\Sigma}_{AC} \\ \mathbf{\Sigma}_{Bu} & \mathbf{\Sigma}_{BA} & \mathbf{\Sigma}_{B} & \mathbf{\Sigma}_{BC} \\ \mathbf{\Sigma}_{Cu} & \mathbf{\Sigma}_{CA} & \mathbf{\Sigma}_{CB} & \mathbf{\Sigma}_{C} \end{pmatrix},$$

and same for $\mathbf{\Lambda}, \mathbf{\Omega}$*. Also, we adopt the following notation:* $\mathbf{\Sigma}^j$ *for* $j = (\boldsymbol{ww}, \boldsymbol{wx}, \boldsymbol{xx})$ *to denote whether the relevant covariance is generated by a cross product or a square product of* $\boldsymbol{W}_{d,i}$ *and* $\boldsymbol{X}_{-GC,i}$*.*
Given these, the following two Lemmas are in order

Lemma 3.1.

$$T^{-1} \sum_{t=1}^{T} \boldsymbol{\tau}_{t,A}^{*\prime} \boldsymbol{\tau}_{t,A}^{*} \xrightarrow{p} \mathbf{\Sigma}_A > 0, \tag{3.33}$$

and

$$\begin{pmatrix} T^{-1/2} \sum_{t=1}^{\lfloor Ts \rfloor} \boldsymbol{u}_t \\ T^{-1/2} \sum_{t=1}^{T} \left(\boldsymbol{\tau}_{t,A}^{*} \otimes \boldsymbol{u}_t \right) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \boldsymbol{B}_u(s) \\ \boldsymbol{\zeta}_1 \end{pmatrix}, \tag{3.34}$$

where $\boldsymbol{B}_u(s)$ *is a vector Brownian motion on* $[0,1]$ *having covariance matrix* $\mathbf{\Omega}_u = \mathbf{\Sigma}_u$ *and* $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \boldsymbol{\zeta}_3)$*, partitioned conformably with* $\boldsymbol{w}_{t,d}^{*}$*, is a normal zero mean random vector with covariance matrix* $\mathbf{\Sigma}_A \otimes \mathbf{\Sigma}_u$ *and* $\boldsymbol{B}_u(s)$*,* $\boldsymbol{\zeta}$ *are independent.*

Lemma 3.2. *The following convergence results of sample moment matrices hold*

(a) $T^{-1/2} \sum_{t=1}^{T} \boldsymbol{u}_t \xrightarrow{d} \boldsymbol{B}_u(1)$

(b) $T^{-1/2} \sum_{t=1}^{T} \boldsymbol{\tau}_{t,A}^{*} \xrightarrow{d} \boldsymbol{B}_A(1)$

(c) $T^{-3/2}\sum_{t=1}^{T}\boldsymbol{\tau}^*_{t,B} \xrightarrow{d} \int_0^1 \boldsymbol{B}_B(s)ds$

(d) $T^{-5/2}\sum_{t=1}^{T}\boldsymbol{\tau}^*_{t,C} \xrightarrow{d} \int_0^1 \bar{\boldsymbol{B}}_C(s)ds$

(e) $T^{-1}\sum_{t=1}^{T}\boldsymbol{\tau}^{*\prime}_{t,B}\boldsymbol{u}_t \xrightarrow{d} \int_0^1 \boldsymbol{B}_B(s)'d\boldsymbol{B}_u(s)$

(f) $T^{-1}\sum_{t=1}^{T}\boldsymbol{\tau}^{*\prime}_{t,B}\boldsymbol{\tau}^*_{t,A} \xrightarrow{d} \int_0^1 \boldsymbol{B}_B(s)'d\boldsymbol{B}_A(s) + \boldsymbol{\Sigma}_{BA} + \boldsymbol{\Lambda}_{BA}$

(g) $T^{-2}\sum_{t=1}^{T}\boldsymbol{\tau}^{*\prime}_{t,B}\boldsymbol{\tau}^*_{t,B} \xrightarrow{d} \int_0^1 \boldsymbol{B}_B(s)'\boldsymbol{B}_B(s)ds$

(h) $T^{-2}\sum_{t=1}^{T}\boldsymbol{\tau}^{*\prime}_{t,C}\boldsymbol{u}_t \xrightarrow{d} \int_0^1 \bar{\boldsymbol{B}}_C(s)'d\boldsymbol{B}_u(s)$

(i) $T^{-2}\sum_{t=1}^{T}\boldsymbol{\tau}^{*\prime}_{t,C}\boldsymbol{\tau}^*_{t,A} \xrightarrow{d} \int_0^1 \bar{\boldsymbol{B}}_C(s)'d\boldsymbol{B}_A(s)$

(l) $T^{-3}\sum_{t=1}^{T}\boldsymbol{\tau}^{*\prime}_{t,C}\boldsymbol{\tau}^*_{t,B} \xrightarrow{d} \int_0^1 \bar{\boldsymbol{B}}_C(s)'\boldsymbol{B}_B(s)ds$

(m) $T^{-4}\sum_{t=1}^{T}\boldsymbol{\tau}^{*\prime}_{t,C}\boldsymbol{\tau}^*_{t,C} \xrightarrow{d} \int_0^1 \bar{\boldsymbol{B}}_C(s)'\bar{\boldsymbol{B}}_C(s)ds,$

where $\bar{\boldsymbol{B}}_C(s) = \int_0^s \boldsymbol{B}_C(u)du.$

Proofs of Lemma 3.3 and Lemma 3.2 are not reported here, we refer to Phillips and Durlauf (1986) and references therein for a full treatment of these results.

Lemma 3.3. *Given below in (3.35) is the usual form of the LM test*

$$LM = \tilde{\boldsymbol{\lambda}}_d'\left[F_d(\hat{\boldsymbol{\phi}}^*)\boldsymbol{S}\left\{\tilde{\boldsymbol{\Sigma}}_{\tilde{\boldsymbol{\mathcal{E}}}}\left(\boldsymbol{W}_d^{*\prime}\mathcal{M}(\boldsymbol{X}_{-GC})\boldsymbol{W}_d^*\right)^{-1}\right\}F_d(\hat{\boldsymbol{\phi}}^*)\boldsymbol{S}'\right]\tilde{\boldsymbol{\lambda}}_d, \tag{3.35}$$

where $\boldsymbol{S}$ is as defined in Section 3.2. This can be rewritten as

$$f_d(\hat{\boldsymbol{\phi}}^*)\boldsymbol{S}' \left[F_d(\hat{\boldsymbol{\phi}}^*)\boldsymbol{S} \left\{ \tilde{\boldsymbol{\Sigma}}_{\tilde{\boldsymbol{\mathcal{E}}}} \left(\boldsymbol{W}_d^{*\prime} \mathcal{M}(\boldsymbol{X}_{-GC}) \boldsymbol{W}_d^* \right)^{-1} \right\} F_d(\hat{\boldsymbol{\phi}}^*)\boldsymbol{S}' \right] f_d(\hat{\boldsymbol{\phi}}^*)\boldsymbol{S}, \tag{3.36}$$

where

$$F_d(\hat{\boldsymbol{\phi}}^*) := \frac{\partial f_d(\hat{\boldsymbol{\phi}}^*)}{\partial \hat{\boldsymbol{\phi}}^*},$$

Restricted LS:

$$\tilde{\boldsymbol{\phi}} = \hat{\boldsymbol{\phi}}^* + \left(\boldsymbol{W}_d^{*\prime} \mathcal{M}(\boldsymbol{X}_{-GC}) \boldsymbol{W}_d^* \right)^{-1} F_d(\hat{\boldsymbol{\phi}}^*)\boldsymbol{S}' \times$$

$$\times \left[F_d(\hat{\boldsymbol{\phi}}^*)\boldsymbol{S} \left(\boldsymbol{W}_d^{*\prime} \mathcal{M}(\boldsymbol{X}_{-GC}) \boldsymbol{W}_d^* \right)^{-1} F_d(\hat{\boldsymbol{\phi}}^*)\boldsymbol{S} \right]^{-1} \times \left[-f_d(\hat{\boldsymbol{\phi}}^*)\boldsymbol{S} \right],$$

$$\tilde{\boldsymbol{\mathcal{E}}} = \left(\boldsymbol{y} - \boldsymbol{W}_d^{*\prime} \tilde{\boldsymbol{\phi}} \right),$$

$$\tilde{\boldsymbol{\Sigma}}_{\tilde{\boldsymbol{\mathcal{E}}}} = (T - K + p)^{-1} \tilde{\boldsymbol{\mathcal{E}}}' \tilde{\boldsymbol{\mathcal{E}}},$$

$$\tilde{\boldsymbol{\lambda}}_d := \left[F_d(\hat{\boldsymbol{\phi}}^*)\boldsymbol{S} \left\{ \tilde{\boldsymbol{\Sigma}}_{\tilde{\boldsymbol{\mathcal{E}}}} \left(\boldsymbol{W}_d^{*\prime} \mathcal{M}(\boldsymbol{X}_{-GC}) \boldsymbol{W}_d^* \right)^{-1} \right\} F_d(\hat{\boldsymbol{\phi}}^*)\boldsymbol{S}') \right]^{-1} f_d(\hat{\boldsymbol{\phi}}^*)\boldsymbol{S}.$$

By pre and post-multiplying by $\boldsymbol{D}_T$ and if we confine our attention to linear Granger (non)-causal hypotheses, we can further rewrite equation (3.36) as

$$LM = (\boldsymbol{D}_T f_d(\hat{\boldsymbol{\phi}}^* - \boldsymbol{\phi}^*)\boldsymbol{S})' \left\{ \tilde{\boldsymbol{\Sigma}}_{\tilde{\mathcal{E}}} \boldsymbol{S} \boldsymbol{D}_T \left(\boldsymbol{W}_d^{*\prime} \mathcal{M}(\boldsymbol{X}_{-GC}) \boldsymbol{W}_d^* \right)^{-1} \boldsymbol{D}_T \boldsymbol{S}' \right\} \times$$

$$\times (\boldsymbol{D}_T f_d(\hat{\boldsymbol{\phi}}^* - \boldsymbol{\phi}^*)\boldsymbol{S}), \tag{3.37}$$

while if we also want to allow for non-linear Granger (non)-causal hypotheses

$$LM = (\boldsymbol{D}_T f_d(\hat{\boldsymbol{\phi}}^* - \boldsymbol{\phi}^*)\boldsymbol{S})' \Bigg\{ (\boldsymbol{D}_T F_d(\hat{\boldsymbol{\phi}}^* - \boldsymbol{\phi}^*)\boldsymbol{S}) \Big(\tilde{\boldsymbol{\Sigma}}_{\tilde{\mathcal{E}}} \boldsymbol{W}_d^{*\prime} \mathcal{M}(\boldsymbol{X}_{-GC}) \boldsymbol{W}_d^* \Big)^{-1} \times \\ \times (\boldsymbol{D}_T F_d(\hat{\boldsymbol{\phi}}^* - \boldsymbol{\phi}^*)\boldsymbol{S})' \Bigg\} (\boldsymbol{D}_T f_d(\hat{\boldsymbol{\phi}}^* - \boldsymbol{\phi}^*)\boldsymbol{S}). \tag{3.38}$$

Appendix B Proof of Lemmas and Theorems in Section 3.4

It suffices for a Gramian matrix to be close in maximum entrywise distance to a matrix which does satisfy the restricted eigenvalue condition, in order to satisfy it as well. The following Lemma 3.4 is the same as Lemma 6 in Kock and Callot (2015) which is in turn directly derived from Lemma 10.1 of Bühlmann and Van De Geer (2011).

Lemma 3.4. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ denote two non-negative definite, r-dimensional square matrices and assume $\boldsymbol{A}$ satisfies the RE condition for some $\boldsymbol{\kappa}_A$. If $\delta = \max_{1\leq i,j\leq r} |A_{ij} - B_{ij}|$, then $\boldsymbol{\kappa}_B^2 \geq \boldsymbol{\kappa}_A^2 - 16s\delta$.*

Proof. Let $\boldsymbol{x} \in \mathbb{R}^r \backslash \{0\}$ and $\forall$ $r \times 1$ vectors $\boldsymbol{x}$ such that $\|\boldsymbol{x}_{\boldsymbol{S}^c}\|_1 \leq 3\|\boldsymbol{x}_{\boldsymbol{S}}\|_1$, then

$$\begin{aligned} \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x} &\leq \left|\boldsymbol{x}'(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{x}\right| \leq ||\boldsymbol{x}||_1 ||(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{x}||_\infty \\ &\leq ||\boldsymbol{x}||_1^2 \delta \leq \delta 16 ||\boldsymbol{x}_{\boldsymbol{S}}||_1^2 \leq \delta 16 s ||\boldsymbol{x}_{\boldsymbol{S}}||^2. \end{aligned}$$

The second and third inequalities follows trivially from application of Holder inequality while the fourth follows by observing that $||\boldsymbol{x}||_1^2\delta = ||\boldsymbol{x}_{\boldsymbol{S}} + \boldsymbol{x}_{\boldsymbol{S}^c}||_1^2\delta \leq ||\boldsymbol{x}_{\boldsymbol{S}} + 3\boldsymbol{x}_{\boldsymbol{S}}||_1^2\delta = 16||\boldsymbol{x}_{\boldsymbol{S}}||_1^2\delta$ where we used the cone condition. Therefore, rearranging

$$\begin{aligned}&\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x} \geq \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} - 16s\delta||\boldsymbol{x}_{\boldsymbol{S}}||^2 \iff \\ &\frac{\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}}{\boldsymbol{x}_{\boldsymbol{S}}'\boldsymbol{x}_{\boldsymbol{S}}} \geq \frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}_{\boldsymbol{S}}'\boldsymbol{x}_{\boldsymbol{S}}} - 16s\delta \geq \boldsymbol{\kappa}_A^2 - 16s\delta.\end{aligned}$$

Hence, by taking the minimum over $\{\boldsymbol{x} \in \mathbb{R}^K \backslash \{0\},\ ||\boldsymbol{x}_{\boldsymbol{S}}||_1 \leq 3||\boldsymbol{x}_{\boldsymbol{S}^c}||_1\}$ proves the statement. □

Before proving the empirical process bound in (c) and Theorem 3.4, a series of preparatory Lemmas is needed:

Lemma 3.5. *Given $i = 1, \dots, K$, let $\epsilon_{i,t} \subset \boldsymbol{\epsilon}$ a zero-mean mds, $\nu_{i,t}$ i.i.d. standard normal sequence, $b_1 \leq \sigma_{ii}^2 \equiv \lim_{T\to\infty} \mathbb{V}ar\left(\sum_{t=1}^T \epsilon_{i,t}\right)/T \leq b_2$ $\forall i$ and b_1, b_2 positive constants, $C \equiv \iota b > 0$, $\iota = 1 + o(V_k/b)$, $V_k = \sum_{k=1}^{[T\ell]} \Psi_L(\nu_k)^2$, $0 < \ell < 1$, $\Psi_L(m) := \exp\left[\frac{\sqrt{1+2Lm}-1}{L}\right]^2 - 1$ $, m \geq 0$, $L > 0$, with ν_k a non-negative sequence.*

For finite K, repeated applications of Lemma 2.2.2 of Van Der Vaart and Wellner (1996) yields

$$\begin{aligned}&\left\|\max_{1\leq i\leq K}\max_{0\leq \ell\leq 1}\left(\left|\sum_{t=1}^{[T\ell]}(\epsilon_{i,t} - \sigma_{ii}\nu_{i,t})\right|\right)\right\|_\Psi \leq \\ &\leq G_1 G_2 \Psi^{-1}(K)\Psi^{-1}(T)\max_{1\leq i\leq K}\max_{0\leq \ell\leq 1}\left\|\left(\left|\sum_{t=1}^{[T\ell]}(\epsilon_{i,t} - \sigma_{ii}\nu_{i,t})\right|\right)\right\|_\Psi,\end{aligned}$$

where G_1, G_2 are constants that depends respectively only on $\Psi^{-1}(K)$, $\Psi^{-1}(T)$. Given the growth of Ψ^{-1} is slowest for rapidly increasing Ψ, then we want to show that the double maximum of the Orlicz norm is bounded. To do so, following Geer and Lederer (2013), we can choose the Bernstein-Orlicz norm entailing functions Ψ_L such that for *each $L > 0$, $m \geq 0$*

$$\begin{aligned} \Psi_L(m) &:= \exp\left[\frac{\sqrt{1+2Lm}-1}{L}\right]^2 - 1, \\ \Psi_L(m)^{-1} &= \sqrt{\log(1+m)} + \frac{L}{2}\log(1+m), \end{aligned} \tag{3.39}$$

then for all $c > \left\|\left(\left|\sum_{t=1}^{[T\ell]}(\epsilon_{i,t} - \sigma_{ii}\nu_{i,t})\right|\right)\right\|_{\Psi_L} \equiv ||Z||_{\Psi_L} =: \tau$ *and by Chebyshev's inequality*

$$\begin{aligned} &\mathbb{P}\left(|Z|/c \geq \sqrt{m} + \frac{Lm}{2}\right) = \mathbb{P}\left(|Z|/c \geq \Psi_L^{-1}(e^m - 1)\right) = \\ &= \mathbb{P}\left(\Psi_L(|Z|/c) \geq e^m - 1\right) \leq \left(\mathbb{E}\Psi_L(|Z|/c) + 1\right) e^{-m}. \end{aligned}$$

Hence, a probability inequality for Z follows as

$$\begin{aligned} \mathbb{P}\left(|Z|/\tau \geq \sqrt{m} + \frac{Lm}{2}\right) &= \lim_{c \downarrow \tau} \mathbb{P}\left(|Z|/c \geq \sqrt{m} + \frac{Lm}{2}\right) \\ &\leq \lim_{c \downarrow \tau} \left(\mathbb{E}\Psi_L(|Z|/c) + 1\right) e^{-m} \leq 2e^{-m}. \end{aligned} \tag{3.40}$$

The tail bound in Lemma 3.5 can be sharpened by looking at the properties of $Z = \left(\sum_{t=1}^{[T\ell]}(\epsilon_{i,t} - \sigma_{ii}\nu_{i,t})\right)$. Note that $\sum_{t=1}^{[T\ell]}\epsilon_{i,t}$ is already by definition of $\boldsymbol{\epsilon}_t$ in Section 3.2 a partial sum of zero-mean mds adapted to the filtration $\mathcal{F}^{\epsilon} = \sigma\left(\epsilon_{i,1}, \ldots, \epsilon_{i,[T\ell]}\right)$ hence it is a martingale by definition. On the other hand, $\sigma_{ii}\sum_{t=1}^{[T\ell]}\nu_{i,t}$ is a martingale sequence w.r.t. its own filtration $\mathcal{F}^{\nu} = \sigma\left(\nu_{i,1}, \ldots, \nu_{i,[T\ell]}\right)$ since it is a partial sum of i.i.d. $N(0,1)$ random variables (cf. zero-mean random walk), hence defining $S_{[T\ell]} := \sum_{k=1}^{[T\ell]}\nu_{i,k}$, then

$$\mathbb{E}[S_{[T\ell]+1}|\mathcal{F}^{\nu}] = \mathbb{E}[\nu_{i,[T\ell]+1} + S_{[T\ell]}|\mathcal{F}^{\nu}] = \mathbb{E}[\nu_{i,[T\ell]+1}] + S_{[T\ell]} = S_{[T\ell]},$$

which shows that $\{S_k\}_{k=1}^{\infty}$ is a martingale sequence. Thus, by telescoping decomposition, $S_{[T\ell]}$ can be rewritten as sum of mds as $S_{[T\ell]} - S_0 =$

$\sum_{k=1}^{[T\ell]} D_k$ for $D_k = S_k - S_{k-1}$, $k \geq 1$.

Let now $S_k^* := \sum_{k=1}^{[T\ell]} (\epsilon_{i,k} - \sigma_{ii}\nu_{i,k})$ be a sequence on $\mathcal{F}_{[T\ell]}^{\Gamma}$ for $\mathcal{F}^{\Gamma} = \mathcal{F}^{\nu} \cup \mathcal{F}^{\epsilon}$. As $\epsilon_{i,t}$ and $\nu_{i,t}$ are independent and mean-zero, then we can show S_k^* is a martingale:

$$\begin{aligned}
&\mathbb{E}\left[S_{[T\ell]+1}^* | \mathcal{F}_{[T\ell]}^{\Gamma}\right] = \\
&\mathbb{E}\left[\left(\epsilon_{i,[T\ell]+1} + \sigma_{ii}\nu_{i,[T\ell]+1}\right) + S_{[T\ell]}^* | \mathcal{F}_{[T\ell]}^{\Gamma}\right] = \\
&\mathbb{E}\left[\left(\epsilon_{i,[T\ell]+1} + \sigma_{ii}\nu_{i,[T\ell]+1}\right)\right] + S_{[T\ell]}^* = \\
&\mathbb{E}\left[\epsilon_{i,[T\ell]+1}\right] + \sigma_{ii}\mathbb{E}\left[\nu_{[T\ell]+1}\right] + S_{[T\ell]}^* = S_{[T\ell]}^*,
\end{aligned}$$

by independence of $S_{[T\ell]+1}^*$ from $\mathcal{F}_{[T\ell]}^{\Gamma}$ and the fact that $S_{[T\ell]}^*$ is measurable on $\mathcal{F}_{[T\ell]}^{\Gamma}$. Then again, by telescoping decomposition we can write S_k^* as partial sum of mds: $S_{[T\ell]}^* - S_0^* = \sum_{k=1}^{[T\ell]} Z_k$ for $Z_k := S_k^* - S_{k-1}^*$, $k \geq 1$ and $\{Z_k\}_{k=1}^{[T\ell]}$ is a mds.

Eq (3.40) implies[15], by equivalent characterization of sub-exponential variables (see e.g. Th.2.13 of Wainwright (2019)), that for $\{(Z_k, \mathcal{F}_k^{\Gamma})\}_{k=1}^{\infty}$ there exists non-negative sequences of numbers (ν_k, α_k) such that $\mathbb{E}[e^{\lambda Z_k} | \mathcal{F}_{k-1}] \leq e^{\frac{\lambda^2 \nu_k^2}{2}}$ for any $|\lambda| < 1/\alpha_k$. Then the following lemma is in order:

Lemma 3.6. *For any $[T\ell]$, $\tau = \left|\left|\left(\sum_{k=1}^{[T\ell]} Z_k\right)\right|\right|_{\Psi} \leq C$ for a constant $C > 0$ and $D_{k,C}$ a positive constant depending only on k and C, a (non-asymptotic) general Bernstein's bound for mds applies such that*

[15] When in (3.40) we use S_k^* in place of Z, then $|S_k^*|$ by standard Jensen's inequality argument is a sub-martingale. Therefore, usual Doob's martingale inequality applies in an analogous way:

$$\mathbb{P}\left(\sup_{0 \leq k \leq [T\ell]} |S_k^*|/\tau \geq \sqrt{m} + \frac{Lm}{2}\right) \leq \lim_{c \downarrow \tau}(\mathbb{E}\boldsymbol{\Psi}_L(|S_k^*|/c) + 1)e^{-m} \leq 2e^{-m}.$$

for $\alpha_* := \max_{k=1,\ldots,[T\ell]} \alpha_k$

$$\mathbb{P}\left(\left|\sum_{k=1}^{[T\ell]} Z_k\right| \geq \tau\left[\sqrt{m}+\frac{Lm}{2}\right]\right) \leq \begin{cases} 2e^{-\frac{(e^m-1)^2}{2D_{k,C}}} & \textit{if } 0 \leq m < \frac{D_{k,C}}{\alpha_*} \\ 2e^{-\frac{(e^m-1)}{\alpha_*}+\frac{m}{2\alpha_*}} & \textit{if } m \geq \frac{D_{k,C}}{\alpha_*}. \end{cases} \tag{3.41}$$

Proof Lemma 3.6. Observe that by conditioning on $\mathcal{F}_{[T\ell]-1}$ and by repeating iterated expectation and sub-exponential definition we have the bound

$$\begin{aligned} \mathbb{E}\left[e^{\lambda\Psi_L\left(\sum_{k=1}^{[T\ell]} Z_k\right)}\right] &\leq \mathbb{E}\left[e^{\lambda\Psi_L\left(\sum_{k=1}^{[T\ell]-1} Z_k\right)}\mathbb{E}\left[e^{\lambda\Psi_L(Z_{[T\ell]})}|\mathcal{F}_{[T\ell]-1}\right]\right] \\ &\leq \mathbb{E}\left[e^{\lambda\Psi_L\left(\sum_{k=1}^{[T\ell]-1} Z_k\right)}\right]e^{\lambda^2\Psi_L(\nu_{[T\ell]})^2/2} \\ &\vdots \\ &\leq e^{\lambda^2\left(\sum_{k=1}^{[T\ell]} \Psi_L(\nu_k)^2\right)/2}, \end{aligned}$$

which shows that $\Psi_L(\sum_{k=1}^{[T\ell]} Z_k)$ is sub-exponential with parameters $(\sqrt{\sum_{k=1}^{[T\ell]} \Psi_L(\nu_k)^2}, \alpha_*)$. Then, to obtain the tail bound in (3.41), by Chernoff-like approach, sub-exponential definition and assuming $\tau < C$ for some positive constant C and given a constant $D_{k,C}$ depending only on k, C, then

$$\begin{aligned} \mathbb{P}\left(\left|\sum_{k=1}^{[T\ell]} Z_k\right|/C \geq \sqrt{m}+\frac{Lm}{2}\right) &= \mathbb{P}\left(\Psi_L\left(\left|\sum_{k=1}^{[T\ell]} Z_k\right|/C\right) \geq e^m - 1\right) \\ &\leq \left(e^{-\lambda(e^m-1)}\mathbb{E}\left[e^{\lambda\Psi_L\left(\left|\sum_{k=1}^{[T\ell]} Z_k\right|/C\right)}\right]\right) \\ &\leq \left(e^{-\lambda(e^m-1)}e^{\lambda^2\left(\sum_{k=1}^{[T\ell]} \Psi_L(\nu_k)^2\right)/C2}\right) \end{aligned}$$

$$= e^{-\lambda(e^m-1)+\lambda^2 D_{k,C}/2} = e^{g(\lambda,m)}.$$

By minimizing the exponent function $g(\lambda, m)$ for λ unconstrained yields $\lambda^* = \frac{e^m-1}{D_{k,C}}$.
If $0 \leq m < \frac{D_{k,C}}{\alpha_*}$ then the unconstrained minimum coincides with the constrained one and
$g^*(m) = -\frac{(e^m-1)^2}{2D_{k,C}}$. For $m \geq \frac{D_{k,C}}{\alpha_*}$, since $g(\cdot, m)$ is monotonically decreasing on $[0, \lambda^*)$ then the constrained minimum is achieved at the boundary $\lambda^\dagger = \alpha_*^{-1}$ such that $g^*(m) = g(\lambda^\dagger, m) = -\frac{e^m+1}{\alpha_*} + \frac{D_{k,C}}{2\alpha_*} \leq -\frac{(e^m-1)}{\alpha_*} + \frac{m}{2\alpha_*}$ which proves the claim. □

Remark 3.14. Lemma 3.6 requires existence of the moment generating function for the sequence
$\Psi_L\left(\left|\sum_{k=1}^{[T\ell]} Z_k\right| /C\right)$ in a neighborhood of zero in order to be able to apply Chernoff's bound. Given Ψ_L a convex, monotone transformation and Z_k a mds composed of a standard Gaussian random variable $\nu_{i,t}$ and a mds $\epsilon_{i,t}$, the only conceived restriction would be on the latter if the true distribution would be e.g., Cauchy or Student-t. In other words, this double-exponentially fast bound is per se invalid for heavy-tailed distribution as the infiniteness of their moment generating functions impede the use of Chernoff's bound. However, if we reduce to the existence of first two moments for $\epsilon_{i,t}$, then the case of Student-t is easily handled: one can in fact rewrite a Student-t as a mixture of zero-mean Gaussian distributions and hence obtain the even moments. As a consequence, one can directly use Chebyshev's inequality as in Geer and Lederer (2013) (see equation (3.40)) to have a (simpler) subexponential bound as $\mathbb{P}\left(\Psi_L\left(\left|\sum_{k=1}^{[T\ell]} Z_k\right| /C\right) \geq e^m - 1\right) \leq 2e^{-m}$. Similarly, for the pathological case of the Cauchy distribution, one can consider the truncated version on a certain large interval in order to obtain finite moments and hence again apply Chebyshev's.

Now we show that the converse result of Lemma 3.5 holds. Namely, given the tail bound in Lemma 3.6 one can obtain a direct bound on the Orlicz norm.

Lemma 3.7. *Given the sub-exponential tail bound in (3.41) and some constants* $\iota = 1 + o(V_k/b)$, b, *such that* $\iota b \equiv C$ *of Lemma 3.6 and* $D_{k,C} = \left(\sum_{k=1}^{[T\ell]} \Psi_L(\nu_k)^2\right)/2C := V_k/2\iota b$, *then*

$$\tau = \left\| \left(\sum_{k=1}^{[T\ell]} Z_k \right) \right\|_{\Psi_{\iota L}} \leq \iota b. \tag{3.42}$$

Proof Lemma 3.7.

$$\begin{aligned}
&\mathbb{E}\Psi_{\iota L}\left(\left|\sum_{k=1}^{[T\ell]} Z_k\right| / \iota b\right) = \int_0^\infty \mathbb{P}\left(\left|\sum_{k=1}^{[T\ell]} Z_k\right| \geq \iota b \Psi_{\iota L}^{-1}(m)\right) dm \\
&= \int_0^\infty \mathbb{P}\left(\left|\sum_{k=1}^{[T\ell]} Z_k\right| \geq \iota b \left[\sqrt{\log(1+m)} + \frac{\iota L}{2}\log(1+m)\right]\right) dm \\
&= \int_0^\infty \mathbb{P}\left(\left|\sum_{k=1}^{[T\ell]} Z_k\right| \geq b \left[\sqrt{\log(1+m)^{\iota^2}} + \frac{\iota L}{2}\log(1+m)^{\iota^2}\right]\right) dm \\
&\leq 2\int_0^\infty \exp\left[-\frac{\left(e^{\log(1+m)^{\iota^2}} - 1\right)^2 \iota b}{2V_k}\right] dm \\
&= 2\int_0^\infty \exp\left[-\frac{\left((1+m)^{\iota^2} - 1\right)^2 \iota b}{2V_k}\right] dm \\
&\overset{u=1+m, du=dm}{=} 2\int_1^\infty \exp\left[-\frac{\left(u^{\iota^2} - 1\right)^2 \iota b}{2V_k}\right] du \\
&\overset{v=u^{\iota^2}-1,\ du=\frac{1}{\iota^2}v^{\frac{1-\iota^2}{\iota^2}}dv}{=} 2\int_0^\infty \exp\left[-\frac{v^2 \iota b}{2V_k}\right] \frac{1}{\iota^2} v^{\frac{1-\iota^2}{\iota^2}} dv
\end{aligned}$$

$$\overset{w=\sqrt{\frac{\iota b}{2V_k}}v,\, dw=\sqrt{\frac{\iota b}{2V_k}}dv}{=} \quad 2\int_0^\infty \exp(-w^2)\frac{1}{\iota^2}\left(\sqrt{\frac{2V_k}{\iota b}}w\right)^{\frac{1-\iota^2}{\iota^2}} dw$$

$$= \frac{2}{\iota^2}\sqrt{\frac{2V_k}{\iota b}}^{\frac{1-\iota^2}{\iota^2}} \int_0^\infty \exp(-w^2)w^{\frac{1}{\iota^2}-1}dw$$

$$\overset{z=w^2\ dw=\frac{1}{2z^{1/2}}dz}{=} \frac{2}{\iota^2}\sqrt{\frac{2V_k}{\iota b}}^{\frac{1}{\iota^2}-1} \int_0^\infty \exp(-z)z^{\frac{1-\iota^2}{2\iota^2}}\frac{1}{2z^{1/2}}dz =$$

$$= \frac{1}{\iota^2}\sqrt{\frac{2V_k}{\iota b}}^{\frac{1}{\iota^2}-1} \int_0^\infty \exp(-z)z^{\frac{1}{2\iota^2}-1}dz$$

$$= \frac{1}{\iota^2}\sqrt{\frac{2V_k}{\iota b}}^{\frac{1}{\iota^2}-1} \Gamma\left(\frac{1}{2\iota^2}\right),$$

and we want the value of the constant ι such that it makes it ≤ 1, for simplicity take the right end of the inequality, such that

$$\Gamma\left(\frac{1}{2\iota^2}\right) = \iota^2\left(\sqrt{\frac{\iota b}{2V_k}}\right)^{\frac{1}{\iota^2}-1}.$$

Because the Γ function has a closed form expression for positive integers, let us assume w.l.o.g. $n := \frac{1}{2\iota^2} \in \mathbb{Z}^+ \iff i = \sqrt{\frac{1}{2n}}$, then by definition:

$$\Gamma(n) = (n-1)! = \left(\frac{1}{2n}\right)\left(\sqrt{\sqrt{\frac{1}{2n}}\frac{b}{2V_k}}\right)^{2n-1} \iff$$

$$n! = \frac{1}{2}\frac{1}{(2n)^{n/2-1/4}}\left(\frac{b}{2V_k}\right)^{n-1/2},$$

and with some straightforward manipulations we can obtain a closed form for the ratio of constants $\frac{b}{V_k}$, namely:

$$\frac{b}{V_k} = 2^{3/2+\frac{1}{n-1/2}}\sqrt{n}(n!)^{\frac{1}{n-1/2}}.$$

For instance, take $\iota = \frac{1}{\sqrt{2}}$, then given $n = 1$, $\frac{b}{V_k} = 2^{7/8}$. Instead, take $\iota = \frac{1}{2}$, then $n = 2$, $\frac{b}{V_k} = 2^{2/3}$. More generally, take $\iota = 1+x$, for $x \to 0$, $a \leq 1$, $V_k/b \to \infty$ then

$$\begin{aligned}
&\frac{1}{\iota^2}\sqrt{\frac{2V_k}{\iota b}}^{\frac{1}{\iota^2}-1}\Gamma\left(\frac{1}{2\iota^2}\right) = a \leq 1\\
&\left(\sqrt{\frac{2V_k}{b}}\right)^{\frac{1}{(1+x)^2}-1}\Gamma\left(\frac{1}{2}\right) = a \leq 1\\
&\left(\sqrt{\frac{2V_k}{b}}\right)^{-2x} = a\frac{1}{\Gamma(1/2)}\\
&-2x\log\left(\sqrt{\frac{2V_k}{b}}\right) = \log\left(\frac{a}{\Gamma(1/2)}\right)\\
&x = \frac{\log\left(\frac{a}{\Gamma(1/2)}\right)}{\log\left(\frac{b}{2V_k}\right)} \to 0^+.
\end{aligned}$$

Hence, again take for simplicity $a = 1$, then

$$\iota = 1 + \frac{\log(\Gamma(1/2))}{\log\left(\frac{2V_k}{b}\right)} + o\left(\frac{V_k}{b}\right),$$

which shows the constant ι converges to 1 and this concludes the proof.

□

We want to allow $K \to \infty$ but the bound in Lemma 3.5 cannot be directly applied to an infinite set. So the problem becomes bounding the expected value of the maximum of an infinite set of random variables.

To handle this and get a bound we need to use a chaining method (cf. Dudley (1978)) as in Theorem 2.2.4 of Van Der Vaart and Wellner (1996). Alternatively, Theorem 8.3 of Cesa-Bianchi and Lugosi (2006) gives a similar bound but under assumption of sub-Gaussian random variables.

Let $(\mathcal{T}, d)$ be an arbitrary semimetric space, for our purpose take the Orlicz space equipped with the semimetric $d(s,t) = \|Z_s - Z_t\|_\psi$. Let $N(\varepsilon, d)$ be the covering number i.e., the minimal number of balls of radius ε needed to cover $\mathcal{T}$. Furthermore, a collection of points is said to be ε -separated if the distance between each pair of points is strictly larger than ε. As a consequence, let $D(\varepsilon, d)$ be the packing number i.e., the maximum number of ε -separated points in $\mathcal{T}$. $\mathcal{T}$ is totally bounded if and only if $N(\varepsilon, d), D(\varepsilon, d) < \infty$, $\forall \varepsilon > 0$ and we assume this is the case. The following proposition is due to Doob (see e.g. Itō and Itåo (2006) Ch. 2.8 for an overview),

Proposition 3.2. *Every real-valued stochastic process defined on a complete probability space has a separable version. Namely, there is a real random process defined on the same probability space which is separable and stocastically equivalent*[16].

Therefore we can assume $Z \equiv \sum_{k=1}^{[T\ell]} Z_k$ to be an almost sure separable stochastic process on the Orlicz space $\mathcal{T}$ i.e., $\sup_{d(s,t)<\delta} |Z_s - Z_t|$ remains constant if $\mathcal{T}$ gets replaced by a suitable subset. Orlicz spaces are Banach space and for our purpose we can assume $\mathcal{T}$ having the strong diameter 2 property[17], namely every finite convex combination of slices

[16] To be precise, the equivalence is in the weak sense of Itō and Itåo (2006), Ch.2.8: two stochastic processes $x_t, t \in T$, and $y_t, t \in T$, are equivalent in the weak sense if

$$P\{\omega / x_t(\omega) = y_t(\omega)\} = 1 \quad \text{for every } t \in T$$

[17] Note that this assumption is reasonable: Orlicz spaces with Luxembourg norm $\left(||Z||_\Psi = \inf\left\{c > 0 : \mathbb{E}\Psi\left(\frac{|Z|}{c} \leq 1\right)\right\}\right)$ can satisfy the diameter 2 property under some conditions on Ψ, see Kamińska et al. (2020) for a recent overview.

of its unit ball $B_{\mathcal{T}}$ has diameter 2. Then, the following Theorem (cf. Van Der Vaart and Wellner (1996), Th. 2.2.4) holds:

Theorem 3.3. *Given Ψ as in (3.39) and given Proposition 3.2 for $Z \equiv \sum_{k=1}^{[T\ell]} Z_k$ such that*

$$||Z_s - Z_t||_{\psi} \leq Cd(s,t), \quad \textit{for every } s,t,$$

for some semimetric d on $\mathcal{T}$ and a constant C, then for any $\eta, \delta > 0$

$$\left\|\sup_{d(s,t)\leq\delta} |Z_s - Z_t|\right\|_{\psi} \leq G\left[\int_0^{\eta} \psi^{-1}(D(\varepsilon,d))d\varepsilon + \delta\psi^{-1}\left(D^2(\eta,d)\right)\right], \tag{3.43}$$

for a constant G depending on ψ and C only.

Proof of Theorem 3.3. The proof follows in the same way as Theorem 2.2.4 in Van Der Vaart and Wellner (1996). First, one constructs nested sets of $\mathcal{T}$ i.e., $\mathcal{T}_0 \subset \mathcal{T}_1 \subset \mathcal{T}_2, \ldots, \subset \mathcal{T}$ such that for every $s,t \in \mathcal{T}_j$, $d(s,t) > \eta 2^{-j}$. As by definition of packing number we have $D(2\varepsilon,d) \leq N(\varepsilon,d) \leq D(\varepsilon,d)$, it follows that the number of points in $\mathcal{T}_j$ will be: $Card(\mathcal{T}_j) \leq D(\eta 2^{-j},d)$. By means of linking every point $t_{j+1} \in \mathcal{T}_{j+1}$ to a unique point $t_j \in \mathcal{T}_j$ we will have that $d(t_j,t_{j+1}) \leq \eta 2^{-j}$. We can therefore create a chain of points from any $t_{k+1} \in \mathcal{T}_{k+1}$ all the way to $t_0 \in \mathcal{T}_0$. Now, for arbitrary points $s_{k+1}, t_{k+1} \in \mathcal{T}_{k+1}$, their increments differences along the chains can be bounded as

$$\begin{aligned}\left|(Z_{s_{k+1}} - Z_{s_0}) - (Z_{t_{k+1}} - Z_{t_0})\right| &= \left|\sum_{j=0}^{k}(Z_{s_{j+1}} - Z_{s_j}) - \sum_{j=0}^{k}(Z_{t_{j+1}} - Z_{t_j})\right| \\ &\leq 2\sum_{j=0}^{k}\max|Z_u - Z_v|,\end{aligned}$$

where if j is fixed, say $j=2$, the maximum is taken over all links (u,v) from $\mathcal{T}_3$ to $\mathcal{T}_2$. It follows that the j-th maximum is taken over at most

$Card(\mathcal{T}_{j+1})$ links i.e., $Card(\mathcal{T}_2)$ continuing the example. Each of such links has bounded Orlicz norm as $||Z_u - Z_v||_\Psi \leq Cd(s,t) \leq \eta 2^{-j}$.

Then we get

$$\left\| \max_{s,t \in \mathcal{T}_{k+1}} |(Z_s - Z_{s_0}) - (Z_t - Z_{t_0})| \right\|_\psi \overset{(i)}{\leq} G \sum_{j=0}^{k} \psi^{-1} \left(D \left(\eta 2^{-j-1}, d \right) \right) \eta 2^{-j}$$
$$\overset{(ii)}{\leq} 4G \int_0^\eta \psi^{-1}(D(\varepsilon, d)) d\varepsilon, \tag{3.44}$$

where (i) follows directly from Lemma 2.2.2 of Van Der Vaart and Wellner (1996), from the fact that $Card(\mathcal{T}_j) \leq D(\eta 2^{-j}, d)$ hence $Card(\mathcal{T}_{j+1}) \leq D(\eta 2^{-j-1}, d)$ and that $\max_{s,t} || \, |(Z_s - Z_{s_0}) - (Z_t - Z_{t_0})| \, ||_\psi$ is bounded by $\eta 2^{-j}$. (ii) follows by extrapolating the extra 2^{-1} from the sum and observing that the map $\eta 2^{-j} = \epsilon \to \psi^{-1}(D(\epsilon, d))$ is non-increasing hence the sum is bounded by the integral.

To conclude the proof, observe that the pointwise increments i.e., $\left| X_{s_{k+1}} - X_{t_{k+1}} \right|$ can be bounded by the left hand side of (3.44) plus the maximum of the discrepancies at the end of the chains: $\max |X_{s_0} - X_{t_0}|$. To analyze the latter, for every chain's endpoint s_0, t_0 starting at two points in $\mathcal{T}_{k+1}$ within δ distance of each other, choose two points $s_{k+1}, t_{k+1} \in \mathcal{T}_{k+1}$ with $d(s_{k+1}, t_{k+1}) < \delta$ and whose chains end in s_0, t_0. By definition of $\mathcal{T}_0$, this gives at most $D^2(\eta, \epsilon)$ pairs and by triangle inequality

$$|X_{s_0} - X_{t_0}| \leq \left| \left(X_{s_0} - X_{s_{k+1}} \right) - \left(X_{t_0} - X_{t_{k+1}} \right) \right| + \left| X_{s_{k+1}} - X_{t_{k+1}} \right|.$$

By taking the maximum and combining previous results we get

$$\| \max_{\substack{s,t \in T_{k+1} \\ d(s,t) < \delta}} |X_s - X_t| \, \|_\psi \leq 8G \int_0^\eta \psi^{-1}(D(\varepsilon, d)) d\varepsilon + \left\| \max \left| X_{s_{k+1}} - X_{t_{k+1}} \right| \right\|_\psi.$$

Here the maximum on the right is taken over the pairs s_{k+1}, t_{k+1} in T_{k+1} uniquely attached to the pairs s_0, t_0 as above. Thus the maximum is over at most $D^2(\eta, d)$ terms, each of whose ψ -norm is bounded by δ. Its ψ-norm is bounded by $K\psi^{-1}\left(D^2(\eta, d)\right)\delta$. Finally by letting $k \to \infty$ gives the claim. □

Corollary 3.1. *Note that given Theorem 3.3 one can obtain a bound on the maximum of the process, such that for any chain endpoint t_0*

$$\left\| \sup_t |Z_t| \right\|_\psi \leq \|Z_{t_0}\|_\psi + G \int_0^{\operatorname{diam} \mathcal{T} \approx 2} \psi^{-1}(D(\varepsilon, d)) d\varepsilon, \tag{3.45}$$

for diam$\mathcal{T}$:= $\sup_{s,t\in\mathcal{T}} d(s,t)$ being the diameter of $\mathcal{T}$ under d, hence a bound on $\|Z_{t_0}\|_\psi$ follows directly from Lemma 3.7. As $D(2\varepsilon, d) \leq N(\varepsilon, d) \leq D(\varepsilon, d)$, then we have

$$\begin{aligned}
&\int_0^2 \sqrt{\log(1 + D(\varepsilon, d))} + \frac{L}{2}\log(1 + D(\varepsilon, d))d\varepsilon \\
\leq &\int_0^2 \sqrt{\log(1 + N(\varepsilon/2, d))} + \frac{L}{2}\log(1 + N(\varepsilon/2, d))d\varepsilon \\
\lessapprox &\int_0^2 \sqrt{\log(2N(\varepsilon/2, d))} + \frac{L}{2}\log(2N(\varepsilon/2, d))d\varepsilon \\
\overset{(i)}{\leq} &\int_0^2 \sqrt{\log(2)} + \sqrt{K\log\left(1 + \frac{4}{\varepsilon}\right)} + \frac{L\log(2)}{2} + \frac{L}{2}\left(K\log\left(1 + \frac{4}{\varepsilon}\right)\right) d\varepsilon \\
= &\int_0^2 \frac{(L+2)\sqrt{K\ln\left(\frac{\varepsilon+4}{\varepsilon}\right)} + \ln(2)\, L + 2\sqrt{\ln(2)}}{2} d\varepsilon < \infty,
\end{aligned} \tag{3.46}$$

where (i) follows from bounding the metric entropy using the volume of the metric unit ball (see e.g. Lemma 5.7 of Wainwright (2019)).

Remark 3.15. Theorem 3.3 bounds the increment of the process Z_t. Every random variable in the supremum is written as sum of small links and the bound depends on the number $(D^2(\eta, d))$ and size (δ) of such links. Z_t is continuous in ψ-Orlicz norm whenever the covering integral converges, which is the case as shown in (3.46). Therefore, the right hand side of (3.43) can be made arbitrarily small by choosing η and δ accordingly.

Theorem 3.4. *Given $i = 1, \ldots, K$, let $\epsilon_{i,t} \subset \boldsymbol{\epsilon}$ a zero-mean mds, $\nu_{i,t}$ i.i.d. standard normal sequence, $b_1 \leq \sigma_{ii}^2 \equiv \lim_{T\to\infty} \mathbb{V}ar\left(\sum_{t=1}^{T} \epsilon_{i,t}\right)/T \leq b_2$ $\forall i$ and b_1, b_2 positive constants, $C \equiv \iota b > 0$, $\iota = 1 + o(V_k/b)$, $V_k = \sum_{k=1}^{[T\ell]} \Psi_L(\nu_k)^2$, $0 < \ell < 1$, $\Psi_L(m) := \exp\left[\frac{\sqrt{1+2Lm}-1}{L}\right]^2 - 1$,$m \geq 0$, $L > 0$, with ν_k a non-negative sequence. Let also, G a constant depending only on Ψ_L, $D(\varepsilon, d)$ be the packing number, in the arbitrary space $\mathcal{T}$ equipped with the semi-metric $d(s,t) = \|Z_s - Z_t\|_{\Psi_L}$ for some separable real-valued stochastic process Z, then the following Gaussian approximation holds:*

$$\begin{aligned}
&\max_{1\leq i\leq K} \max_{0\leq \ell\leq 1} \mathbb{E}\left[\left|\sum_{t=1}^{[T\ell]} (\epsilon_{i,t} - \sigma_{ii}\nu_{i,t})\right|\right] = \\
&= \mathcal{O}\left(C + G\int_0^{diam\mathcal{T}} \sqrt{\log(1 + D(\varepsilon, d))} + \frac{L}{2}\log(1 + D(\varepsilon, d))d\varepsilon\right) \qquad (3.47)\\
&=: \bar{\gamma}_T.
\end{aligned}$$

Remark 3.16. The provided uniform bound to the empirical process implies that the choice of the tuning parameter λ reduces to majorate the quantity on the right hand side i.e., for $\nu > 1$, $\lambda > \nu 3\bar{\gamma}_T$, (see e.g. Lederer and Vogt, 2020). The role of the tuning parameter λ is in fact to introduce more bias towards zero to reduce the noise. The constant ν appears in the lasso consistency rate as shown later in the verification of Assumption 4, (e). It is worth noting that we are not assuming Gaussian errors nor approximating the distribution of only $\boldsymbol{u}$ to be Gaussian as typically done in the literature. In fact, assuming Gaussian errors one can easily find that a crude but working choice for the tuning parameter would be $\lambda \geq \sqrt{\log K/T}$. Our approach is more general as it requires $\boldsymbol{\epsilon}$ only to be martingale difference and as such the whole empirical process is approximately Gaussian.

From Lemma 3.1 and results (e),(h) of Lemma 3.2 in Appendix A, the following standard convergence results for the three separate blocks of $\boldsymbol{X}'\boldsymbol{u}$ follows:

$$\boldsymbol{D}_T^{-1}\left[\boldsymbol{X}_A'\boldsymbol{u}^{(j)},\boldsymbol{X}_B'\boldsymbol{u}^{(j)},\boldsymbol{X}_C'\boldsymbol{u}^{(j)}\right]'\xrightarrow{d}\left[\boldsymbol{\zeta}_1,\int_0^1\boldsymbol{B}(s)d\boldsymbol{B}_{\boldsymbol{u}}(s)',\int_0^1\bar{\boldsymbol{B}}(s)d\boldsymbol{B}_{\boldsymbol{u}}(s)'\right]',$$

where $\boldsymbol{B}(s)$ represents a vector Brownian motion. The following theorem uses the bound in Theorem 3.4 to obtain a Gaussian approximation for each sub-parts of $\boldsymbol{X}'\boldsymbol{u}$, corresponding to $I(0)$, $I(1)$, $I(2)$ variables in $\boldsymbol{X}$.

Proof of Theorem 3.4. By Jensen's inequality

$$\max_{1\le i\le K}\max_{0\le \ell\le 1}\mathbb{E}\left(\left|\sum_{t=1}^{[T\ell]}(\epsilon_{i,t}-\sigma_{ii}\nu_{i,t})\right|\right)\le\mathbb{E}\max_{1\le i\le K}\max_{0\le \ell\le 1}\left(\left|\sum_{t=1}^{[T\ell]}(\epsilon_{i,t}-\sigma_{ii}\nu_{i,t})\right|\right),$$

now let us consider the Orlicz norm: $||Z||_\Psi := \inf\{c>0:\mathbb{E}\Psi\left(\frac{|Z|}{c}\right)\le 1\}$ of the right hand side, consider a non-decreasing, convex function $\Psi:[0,\infty)\to[0,\infty)$ such that $\Psi(0)=0$ and a constant $C>0$ such that

$$\inf\left\{C>0:\mathbb{E}\Psi\left(\left|\mathbb{E}\max_{1\le i\le K}\max_{0\le \ell\le 1}\left(\left|\sum_{t=1}^{[T\ell]}(\epsilon_{i,t}-\sigma_{ii}\nu_{i,t})\right|\right)\right|/C\right)\le 1\right\},\tag{3.48}$$

note that using $|\mathbb{E}(X)|\le\mathbb{E}(|X|)$ and again Jensen's inequality

$$\mathbb{E}\Psi\left(\left|\mathbb{E}\max_{1\le i\le K}\max_{0\le \ell\le 1}\left(\left|\sum_{t=1}^{[T\ell]}(\epsilon_{i,t}-\sigma_{ii}\nu_{i,t})\right|\right)\right|/C\right)\le$$
$$\le\mathbb{E}\left(\mathbb{E}\left(\Psi\left(\left|\max_{1\le i\le K}\max_{0\le \ell\le 1}\left(\left|\sum_{t=1}^{[T\ell]}(\epsilon_{i,t}-\sigma_{ii}\nu_{i,t})\right|\right)\right|/C\right)\right)\right),$$

so we can drop an expectation and (3.48) assumes the usual Orlicz form:

$$\left\| \max_{1\le i\le K} \max_{0\le \ell\le 1} \left(\left| \sum_{t=1}^{[T\ell]} (\epsilon_{i,t} - \sigma_{ii}\nu_{i,t}) \right| \right) \right\|_{\Psi}.$$

For finite K, repeted applications of Lemma 2.2.2 of Van Der Vaart and Wellner (1996) yields

$$\begin{aligned} &\left\| \max_{1\le i\le K} \max_{0\le \ell\le 1} \left(\left| \sum_{t=1}^{[T\ell]} (\epsilon_{i,t} - \sigma_{ii}\nu_{i,t}) \right| \right) \right\|_{\Psi} \le \\ &\le G_1 G_2 \Psi^{-1}(K)\Psi^{-1}(T) \max_{1\le i\le K} \max_{0\le \ell\le 1} \left\| \left(\left| \sum_{t=1}^{[T\ell]} (\epsilon_{i,t} - \sigma_{ii}\nu_{i,t}) \right| \right) \right\|_{\Psi}. \end{aligned} \tag{3.49}$$

Given results in Corollary 3.1, (3.49) is extended to allow $K \to \infty$. Then, equation (3.45) for Ψ_L defined as in Lemma 3.5 coupled with results of Corollary 3.1, Lemma 3.5, 3.6, 3.7 give the final expression of the bound and conclude the proof. □

Theorem 3.5. *Given results in Theorem 3.4, the empirical process bound in Assumption 4, (c) is formalized as*

$$\begin{aligned} &\boldsymbol{D}_T^{-1}\left[\boldsymbol{X}_A'\boldsymbol{u}^{(j)}, \boldsymbol{X}_B'\boldsymbol{u}^{(j)}, \boldsymbol{X}_C'\boldsymbol{u}^{(j)}\right]' - \left[\boldsymbol{\zeta}_1, \int_0^1 \boldsymbol{B}(s)d\boldsymbol{B}_{\boldsymbol{u}}(s)', \int_0^1 \bar{\boldsymbol{B}}(s)d\boldsymbol{B}_{\boldsymbol{u}}(s)'\right]' = \\ &= \left[\mathcal{O}(T^{-1/2}), \mathcal{O}(A \wedge B), 2\mathcal{O}(A)\right] \end{aligned} \tag{3.50}$$

where $\mathcal{O}(A) = \mathcal{O}\left(C + G\int_0^{diam\mathcal{T}} \sqrt{\log(1 + D(\varepsilon, d))} + \frac{L}{2}\log(1 + D(\varepsilon, d))d\varepsilon\right)$ *as in Theorem 3.4,* $\mathcal{O}(B) = \mathcal{O}\left(\sqrt{\frac{\log(T)}{T}}\right)$.

Proof of Theorem 3.5. The first term, corresponding to the stationary part of $\boldsymbol{X}$, follows straightforwardly from standard CLT results:

$$|T^{-1/2}\boldsymbol{X}'_{\boldsymbol{A}}\boldsymbol{u}^{(j)} - \zeta_1| = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

The second block of $\boldsymbol{X}$ corresponds to $I(1)$ variables, hence we have

$$\begin{aligned}
&T^{-1}\boldsymbol{X}'_{\boldsymbol{B}}\boldsymbol{u}^{(j)} - \int_0^1 \boldsymbol{B}(s)d\boldsymbol{B}_{\boldsymbol{u}}(s)' \\
\equiv&T^{-1}\sum_{t=1}^{T} X^B_{i,t}u_{i,t} - \int_0^1 \boldsymbol{B}_i(s)d\boldsymbol{B}^{\boldsymbol{u}}_i(s)' \\
=&\sum_{t=1}^{T}\left(T^{-1/2}X^B_{i,t}\left(T^{-1/2}u_{i,t} - \int_{\frac{t-1}{T}}^{\frac{t}{T}} d\boldsymbol{B}^{\boldsymbol{u}}_i(s)\right) + \right. \\
&\left. + \int_{\frac{t-1}{T}}^{\frac{t}{T}} \left[T^{-1/2}X^B_{i,t} - \boldsymbol{B}_i(s)\right] d\boldsymbol{B}^{\boldsymbol{u}}_i(s)\right) \qquad (3.51) \\
=&\sum_{t=1}^{T}\left(T^{-1/2}X^B_{i,t}\underbrace{\left(T^{-1/2}u_{i,t} - \boldsymbol{B}_i(T^{-1})\right)}_{:=w_t} + \right. \\
&\left. + \int_{\frac{t-1}{T}}^{\frac{t}{T}} \left[T^{-1/2}X^B_{i,t} - \boldsymbol{B}_i(s)\right] d\boldsymbol{B}^{\boldsymbol{u}}_i(s)\right).
\end{aligned}$$

In order to take care of the left hand side of (3.51), given $W_t := \sum_{s=1}^{t} w_t$:

$$\begin{aligned}
\sum_{t=1}^{T} T^{-1/2}X^B_{i,t}w_t &= \sum_{t=1}^{T}\left(T^{-1/2}X^B_{i,t}\left(T^{-1/2}u_{i,t} - \boldsymbol{B}^{\boldsymbol{u}}_i(T^{-1})\right)\right) \\
&= T^{-1/2}X^B_{i,t}W_T - T^{-1/2}\sum_{t=1}^{T-1} u_{i,t}W_t,
\end{aligned}$$

and by Theorem 3.4, the terms $W_t = \sum_{t=1}^{T} T^{-1/2}u_{i,t} - \boldsymbol{B}_i(T^{-1})$ can be bounded as by Markov's inequality $\mathbb{P}(W_t > \gamma) \leq \frac{\mathbb{E}W_t}{\gamma}$ and upon taking the maxima:

$$\max_{1\le i\le K}\max_{0\le \ell\le 1}\gamma^{-1}\mathbb{E}\left[\left|\sum_{t=1}^{[T\ell]}(u_{i,t}-\boldsymbol{B}(\ell))\right|\right]=$$
$$=\mathcal{O}\left(C+G\int_0^{diam\mathcal{T}}\sqrt{\log(1+D(\varepsilon,d))}+\frac{L}{2}\log(1+D(\varepsilon,d))d\varepsilon\right).$$

The right hand side of (3.51) can be handled by deriving an upper bound for its integrand, namely:

$$\sup_{t\le T,\ \frac{t-1}{T}\le s\le\frac{t}{T}}\left|T^{-1/2}X_{i,t}^{B}-\boldsymbol{B}_i(s)\right|$$
$$\le\underbrace{\sup_{t\le T}\left|T^{-1/2}X_{i,t}^{B}-\boldsymbol{B}_i\left(\frac{t-1}{T}\right)\right|}_{(a)}+\underbrace{\sup_{\frac{t-1}{T}\le s\le\frac{t}{T}}\left|\boldsymbol{B}_i(s)-\boldsymbol{B}_i\left(\frac{t-1}{T}\right)\right|}_{(b)}. \tag{3.52}$$

To bound (a) we can use again our Theorem 3.4. In fact, $X_{i,t}^{B}$ can be rewritten as partial sum process of i.i.d. sequence $\{u_t\}$ thus as above by Markov's inequality we have

$$\sup_{t\le T}\left|T^{-1/2}X_{i,t}^{B}-\boldsymbol{B}_i\left(\frac{t-1}{T}\right)\right|\le\max_{1\le i\le K}\max_{0\le \ell\le 1}\gamma^{-1}\mathbb{E}\left[\left|\sum_{t=1}^{[T\ell]}(u_{i,t}-\boldsymbol{B}_i(\ell))\right|\right]$$
$$=\mathcal{O}\left(C+G\int_0^{diam\mathcal{T}}\sqrt{\log(1+D(\varepsilon,d))}+\frac{L}{2}\log(1+D(\varepsilon,d))d\varepsilon\right).$$

Bounding (b) follows from Levy's modulus of continuity theorem, namely, let $h=s-\frac{t-1}{T}$

$$\lim_{h\downarrow 0}\sup_{\frac{t-1}{T}\le s-h\le\frac{t}{T}}\frac{\left|B(\frac{t-1}{T}+h)-B(\frac{t-1}{T})\right|}{\sqrt{h\log\left(\frac{1}{h}\right)}}=1,$$

hence

$$(b) \leq \sqrt{h \log\left(\frac{1}{h}\right)} = \mathcal{O}\left(\sqrt{\frac{\log(T)}{T}}\right).$$

The third and last block of $\boldsymbol{X}$ corresponds to $I(2)$ variables, hence we have

$$\begin{aligned}
&T^{-2}\boldsymbol{X}_{\boldsymbol{C}}'\boldsymbol{u}^{(j)} - \int_0^1 \bar{\boldsymbol{B}}(s)d\boldsymbol{B}_{\boldsymbol{u}}(s)' \\
\equiv &T^{-2}\sum_{t=1}^{T} X_{i,t}^{C}u_{i,t} - \int_0^1 \left(\int_0^s \boldsymbol{B}_i(u)du\right) d\boldsymbol{B}_i^{\boldsymbol{u}}(s)' \\
= &\sum_{t=1}^{T}\left(T^{-1}X_{i,t}^{C}\left(T^{-1}u_{i,t} - \int_{\frac{t-1}{T}}^{\frac{t}{T}} d\boldsymbol{B}_i^{\boldsymbol{u}}(s)\right) + \right. \qquad (3.53)\\
&\left. + \int_{\frac{t-1}{T}}^{\frac{t}{T}} \left[T^{-1}X_{i,t}^{C} - \left(\int_0^s \boldsymbol{B}_i(u)du\right)\right] d\boldsymbol{B}_i^{\boldsymbol{u}}(s)\right) \\
= &\sum_{t=1}^{T}\left(T^{-1}X_{i,t}^{C}\underbrace{\left(T^{-1}u_{i,t} - \boldsymbol{B}_i(T^{-1})\right)}_{:=w_t} + \right. \\
&\left. + \int_{\frac{t-1}{T}}^{\frac{t}{T}} \left[T^{-1}X_{i,t}^{C} - \left(\int_0^s \boldsymbol{B}_i(u)du\right)\right] d\boldsymbol{B}_i^{\boldsymbol{u}}(s)\right).
\end{aligned}$$

The left hand side in (3.53) follows in the same exact way as in (3.51). The right hand side, by rewriting the $I(2)$ process as double sum of innovations:

$$\int_{\frac{t-1}{T}}^{\frac{t}{T}} \left[T^{-1}X_{i,t}^{C} - \left(\int_0^s \boldsymbol{B}_i(u)du\right)\right] d\boldsymbol{B}_i^{\boldsymbol{u}}(s)$$

$$\int_{\frac{t-1}{T}}^{\frac{t}{T}} \left(\underbrace{T^{-1}\sum_{k=1}^{t}\sum_{j=1}^{k} u_{i,j} - \sum_{k=1}^{t}\int_{\frac{(k-1)s}{t}}^{\frac{k}{t}s} \boldsymbol{B}_i(u)du}_{(a)}\right) d\boldsymbol{B}_i^{\boldsymbol{u}}(s).$$

From Theorem 3.4 it follows immediately that by taking the supremum over the outer integral limits of the integrand that

$$\sup_{t\leq T,\ \frac{t-1}{T}\leq u\leq\frac{t}{T}} |(a)| \leq \max_{1\leq i\leq K}\max_{0\leq\ell\leq 1}\gamma^{-1}\mathbb{E}\left[\left|\sum_{k=1}^{[t\ell]}\left(\sum_{j=1}^{k} u_{i,j} - \boldsymbol{B}_i(\ell)\right)\right|\right]$$
$$= \mathcal{O}_p\left(C + G\int_0^{diam\mathcal{T}} \sqrt{\log(1+D(\varepsilon,d))} + \frac{L}{2}\log(1+D(\varepsilon,d))d\varepsilon\right).$$

which concludes the proof.

□

Verification of Assumption 4, (e). Let $\boldsymbol{X}^{\dagger} = \boldsymbol{X}\boldsymbol{D}_T^{-1}$, $\hat{\boldsymbol{\beta}}^{\dagger} = \boldsymbol{D}_T\hat{\boldsymbol{\beta}}$ and likewise for $\boldsymbol{\beta}^{\dagger(0)}$. Recall from (f) that $\boldsymbol{S}_0$ is the true support and $\boldsymbol{S}_0^c$ is its complement

$$(2T)^{-1}\left\|\boldsymbol{y} - \boldsymbol{X}^{\dagger}\hat{\boldsymbol{\beta}}\right\|_2^2 + \lambda\left\|\hat{\boldsymbol{\beta}}^{\dagger}\right\|_1 \leq (2T)^{-1}\left\|\boldsymbol{y} - \boldsymbol{X}^{\dagger}\boldsymbol{\beta}^{\dagger(0)}\right\|_2^2 + \lambda\left\|\boldsymbol{\beta}^{\dagger(0)}\right\|_1,$$
$$(2T)^{-1}\left\|\boldsymbol{X}^{\dagger}(\hat{\boldsymbol{\beta}}^{\dagger} - \boldsymbol{\beta}^{\dagger(0)})\right\|_2^2 - T^{-1}\left\|\boldsymbol{u}'\boldsymbol{X}^{\dagger}(\hat{\boldsymbol{\beta}}^{\dagger} - \boldsymbol{\beta}^{\dagger(0)})\right\|_2^2 + \lambda\left\|\hat{\boldsymbol{\beta}}^{\dagger}\right\|_1 \leq \lambda\left\|\boldsymbol{\beta}^{\dagger(0)}\right\|_1,$$
$$(2T)^{-1}\left\|\boldsymbol{X}^{\dagger}(\hat{\boldsymbol{\beta}}^{\dagger} - \boldsymbol{\beta}^{\dagger(0)})\right\|_2^2 - T^{-1}\left\|\boldsymbol{X}^{\dagger\prime}\boldsymbol{u}\right\|_{\infty}\left\|(\hat{\boldsymbol{\beta}}^{\dagger} - \boldsymbol{\beta}^{\dagger(0)})\right\|_1 \leq \lambda\left\|\boldsymbol{\beta}^{\dagger(0)}\right\|_1 - \lambda\left\|\hat{\boldsymbol{\beta}}^{\dagger}\right\|_1,$$
$$(2T)^{-1}\left\|\boldsymbol{X}^{\dagger}(\hat{\boldsymbol{\beta}}^{\dagger} - \boldsymbol{\beta}^{\dagger(0)})\right\|_2^2 - \frac{\lambda}{\nu}\left\|(\hat{\boldsymbol{\beta}}^{\dagger} - \boldsymbol{\beta}^{\dagger(0)})\right\|_1 \leq \lambda\left\|(\hat{\boldsymbol{\beta}}^{\dagger}_{\boldsymbol{S}_0} - \boldsymbol{\beta}^{\dagger(0)})\right\|_1 - \lambda\left\|\hat{\boldsymbol{\beta}}^{\dagger}_{\boldsymbol{S}_0^c}\right\|_1,$$
$$(2T)^{-1}\left\|\boldsymbol{X}^{\dagger}(\hat{\boldsymbol{\beta}}^{\dagger} - \boldsymbol{\beta}^{\dagger(0)})\right\|_2^2 - \frac{\lambda}{\nu}\left\|(\hat{\boldsymbol{\beta}}^{\dagger}_{\boldsymbol{S}_0} - \boldsymbol{\beta}^{\dagger(0)})\right\|_1 - \frac{\lambda}{\nu}\left\|\hat{\boldsymbol{\beta}}^{\dagger}_{\boldsymbol{S}_0^c}\right\|_1 \leq$$

$$
\begin{aligned}
&\leq \lambda \left\| (\hat{\boldsymbol{\beta}}^{\dagger}_{\boldsymbol{S}_0} - \boldsymbol{\beta}^{\dagger(0)}) \right\|_1 - \lambda \left\| \hat{\boldsymbol{\beta}}^{\dagger}_{\boldsymbol{S}_0^c} \right\|_1, \\
&(2T)^{-1} \left\| \boldsymbol{X}^{\dagger}(\hat{\boldsymbol{\beta}}^{\dagger} - \boldsymbol{\beta}^{\dagger(0)}) \right\|_2^2 + \lambda \left(1 - \frac{1}{\nu}\right) \left\| \hat{\boldsymbol{\beta}}^{\dagger}_{\boldsymbol{S}_0^c} \right\|_1 \leq \lambda \left(1 + \frac{1}{\nu}\right) \left\| (\hat{\boldsymbol{\beta}}^{\dagger}_{\boldsymbol{S}_0} - \boldsymbol{\beta}^{\dagger(0)}) \right\|_1, \\
&(2T)^{-1} \left\| \boldsymbol{X}^{\dagger}(\hat{\boldsymbol{\beta}}^{\dagger} - \boldsymbol{\beta}^{\dagger(0)}) \right\|_2^2 \leq \lambda \left(1 + \frac{1}{\nu}\right) \sqrt{\bar{s}_T} \left\| (\hat{\boldsymbol{\beta}}^{\dagger}_{\boldsymbol{S}_0} - \boldsymbol{\beta}^{\dagger(0)}) \right\|_2 \\
&\leq \lambda \left(1 + \frac{1}{\nu}\right) \sqrt{\bar{s}_T} \boldsymbol{\kappa}_{T,\min}^{-1} \left\| \boldsymbol{X}^{\dagger}(\hat{\boldsymbol{\beta}}^{\dagger} - \boldsymbol{\beta}^{\dagger(0)}) \right\|_2, \\
&T^{-1} \left\| \boldsymbol{X}^{\dagger}(\hat{\boldsymbol{\beta}}^{\dagger} - \boldsymbol{\beta}^{\dagger(0)}) \right\|_2 \leq 2 \left(1 + \frac{1}{\nu}\right) \frac{\lambda \sqrt{\bar{s}_T}}{\boldsymbol{\kappa}_{T,\min}}, \\
&T^{-1} \left\| \boldsymbol{X} \boldsymbol{D}_T^{-1} (\boldsymbol{D}_T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)}) \right\|_2 \leq 2 \left(1 + \frac{1}{\nu}\right) \frac{\lambda \sqrt{\bar{s}_T}}{\boldsymbol{\kappa}_{T,\min}}, \\
&T^{-1} \left\| \boldsymbol{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)}) \right\|_2 \leq 2 \left(1 + \frac{1}{\nu}\right) \frac{\lambda \sqrt{\bar{s}_T}}{\boldsymbol{\kappa}_{T,\min}},
\end{aligned}
$$

and taking the square of both sides gives the claim. □

Verification of Assumption 4, (g). (I) entails stationary variables and as such it is well known to hold (see e.g. Kock and Callot (2015) and Medeiros and Mendes (2016a) for, respectively, Gaussian and non-Gaussian innovations) whenever the minimum eigenvalue of the corresponding covariance matrix is bounded away from zero. To see this, consider writing the specific (positive-definite) covariance (sub)-matrix as $\boldsymbol{\Sigma}_{A_{\hat{s}}} = \mathbb{E}(\boldsymbol{X}_{A_{\hat{s}}} \boldsymbol{X}'_{A_{\hat{s}}})$ and let $0 < \boldsymbol{\kappa}_{\hat{\boldsymbol{\Sigma}}} < 1$ be a constant, then it is possible to bound the maximal absolute difference between the Gram matrix and the covariance as:

$$
\begin{aligned}
&\mathbb{P}\left(\max_{1 \leq i,j \leq s_{A_{\hat{s}}}} \left[|\hat{\boldsymbol{\Sigma}}_{A_{\hat{s}}} - \boldsymbol{\Sigma}_{A_{\hat{s}}}| \right]_{i,j} \geq \frac{\boldsymbol{\kappa}_{\hat{\boldsymbol{\Sigma}}}}{s_A} \right) \\
&\leq {s_{A_{\hat{s}}}}^2 \max_{1 \leq i,j \leq \boldsymbol{s_A}} \mathbb{P}\left(\left[|\hat{\boldsymbol{\Sigma}}_{A_{\hat{s}}} - \boldsymbol{\Sigma}_{A_{\hat{s}}}| \right]_{i,j} \geq \frac{T \boldsymbol{\kappa}_{\hat{\boldsymbol{\Sigma}}}}{s_{A_{\hat{s}}}} \right) \\
&\leq 2 c_1 s^2 T^{\gamma_1/2} \exp\left[- \frac{c_2 \left(\phi_{\min}/s\right)^2 T^{1-\gamma_1-\gamma_2}}{288} \right] + c_3 \frac{s^3}{\phi_{\min}} \boldsymbol{D}_{k,T} + \frac{s^3}{\phi_{\min}} \boldsymbol{E}_T,
\end{aligned}
$$

where the first step follows from the union bound and the second from the Triplex inequality of Jiang (2009) with $c_{1,2,3}, \gamma_{1,2} > 0$. The dependence term is $\boldsymbol{D}_{k,T} = T^{-1}\mathbb{E}\left|\mathbb{E}\left(\boldsymbol{X}'_{A_{\hat{s}}}\boldsymbol{X}_{A_{\hat{s}}}|\mathcal{F}_{t-m}\right) - \mathbb{E}\boldsymbol{X}'_{A_{\hat{s}}}\boldsymbol{X}_{A_{\hat{s}}}\right|$ where $\{\mathcal{F}_t\}_{t=-\infty}^{\infty}$ is an increasing sequence of σ-fields and m a positive integer defining the dependence window. The tail term is $\boldsymbol{E}_T = T^{-1}\mathbb{E}\left|\boldsymbol{X}'_{A_{\hat{s}}}\boldsymbol{X}_{A_{\hat{s}}}\right| I\left(\left|\boldsymbol{X}'_{A_{\hat{s}}}\boldsymbol{X}_{A_{\hat{s}}}\right| > C\right)$. Since $\boldsymbol{X}_{A_{\hat{s}}}$ is $I(0)$, it can be written as a VMA(∞) process as $\boldsymbol{X}_{A_{\hat{s}}} = (\boldsymbol{x}_{1,A}, \boldsymbol{x}_{2,A} \dots, \boldsymbol{x}_{\boldsymbol{s_A},A})' = \sum_{j=0}^{\infty}\boldsymbol{\theta}_j\boldsymbol{u}_{t-j}$ with $\{\boldsymbol{u}_t, \mathcal{F}_{u,t-1}\}_{-\infty}^{\infty}$ being a martingale difference sequence with respect to an increasing sequence of σ-fields $\mathcal{F}_{u,t-1} = \sigma\left\{\boldsymbol{u}_{t-1}, \boldsymbol{u}_{t-2}, \dots\right\}$ such that $\mathbb{E}\left[\boldsymbol{u}_t|\mathcal{F}_{u,t-1}\right] = 0$, $\mathbb{E}\left[\boldsymbol{u}_t\boldsymbol{u}'_t|\mathcal{F}_{u,t-1}\right] = \boldsymbol{\Sigma}_u$ then it is not restrictive to assume the moment boundedness of the marginals $\mathbb{E}\left|\boldsymbol{x}_{i,A}\right|^p < c_{i,p}$ $\forall i = 1, \dots, s_A$, for $c_p > 0$, $p > 1$ as long as $|\boldsymbol{\theta}|_1 < \infty$ i.e., the absolute summability of the coefficients and $\mathbb{E}\left|\boldsymbol{u}_r\right|^p < \infty, \forall r$ finite moments are also assumed.

Then, $\forall i = 1, \dots, s_A$ it follows by (1) Holder and (2) Markov's inequalities

$$\begin{aligned}\mathbb{E}\left[\left|\boldsymbol{x}_{i,A}\right| I\left(\left|\boldsymbol{x}_{i,A}\right| > C\right)\right] &\overset{(1)}{\leq} \mathbb{E}\left(\left|\boldsymbol{x}_{i,A}\right|^p\right)^{\frac{1}{p}} \mathbb{P}\left(\left|\boldsymbol{x}_{i,A}\right| > C\right)^{\frac{(p-1)}{p}} \\ &\overset{(2)}{\leq} \frac{\mathbb{E}\left(\left|\boldsymbol{x}_{i,A}\right|^p\right)^{\frac{1}{p}} \mathbb{E}\left(\left|\boldsymbol{x}_{i,A}\right|^p\right)^{\frac{(p-1)}{p}}}{C^{\frac{p(p-1)}{p}}} = \mathbb{E}\left(\left|\boldsymbol{x}_{i,A}\right|^p\right) C^{-(p-1)}\end{aligned}$$

Hence, for $\mathbb{E}|\boldsymbol{u}|^{2p} < c_{i,2p}$, then the "tail" term $\boldsymbol{E}_T$ by Cauchy-Schwarz is bounded as:

$$\boldsymbol{E}_T \leq \sqrt{\prod_{i=1}^{\boldsymbol{s_A}} c_{i,2p} C^{-(p-1)}}.$$

The "dependence" term $\boldsymbol{D}_{k,T}$ can be bounded using mixingales arguments (see Davidson (1994) Ch.16). The linear VMA process $\boldsymbol{X}_{A_{\hat{s}}} = \sum_{j=0}^{\infty}\boldsymbol{\theta}_j\boldsymbol{u}_{t-j}$, where $\boldsymbol{u}_t$ is an L_p-bounded martingale difference ($p \geq 1$) and given the absolute summability of the coefficients, it is in fact a mixingale process i.e., $\left\|\mathbb{E}\left(\boldsymbol{X}_{A_{\hat{s}}}|\mathcal{F}_{t-m}\right)\right\|_p \leq c_t\zeta_m$ where by Minkowski inequality $c_t = \sup_r \|\boldsymbol{u}_s\|$ is a non-negative constant and coefficients $\zeta_m = \sum_{j=m}^{\infty}|\theta_{i,j}| \to 0$ as $m \to \infty$ by the absolute summability of the coefficients. Hence, by mixingales definition the "dependence" term is bounded as $\boldsymbol{D}_{k,T} \leq c_t\zeta_m$. Finally, by application of Lemma 3.4 (from

Lemma 6.17 in Bühlmann and Van De Geer (2011)) we obtain $\inf_{\boldsymbol{x}'\boldsymbol{x}=1} \boldsymbol{x}'\widehat{\boldsymbol{\Sigma}}_{A_{\hat{s}}}\boldsymbol{x} > \inf_{\boldsymbol{x}'\boldsymbol{x}=1} \boldsymbol{x}'\boldsymbol{\Sigma}_{A_{\hat{s}}}\boldsymbol{x} - s\frac{\boldsymbol{\kappa}_{\hat{\boldsymbol{\Sigma}}}}{s} > \boldsymbol{\kappa}_{\hat{\boldsymbol{\Sigma}}}$ which concludes the proof. □

For the terms (II), (III) i.e., for $I(1)$ and $I(2)$ variables the verification of the restricted eigenvalue condition is more complex as their scaled Gram matrices do not converge in probability to a full-rank matrix but instead they converge weakly to matrix stochastic integrals (see Lemma 3.2). Smeekes and Wijler (2021) showed how by allowing for a factor $\boldsymbol{s}^2$ to multiply the Gram matrix, then $\boldsymbol{\kappa}^2_{\boldsymbol{\Sigma}}(\boldsymbol{s}) > 0$ on a set with probability converging to one in the unit root non stationary case. To the best of our knowledge this is the first result in the literature extending the restricted eigenvalue condition to unit root non-stationary data.

Appendix C Main results on PDS-LA-LM Test

Proof of Theorem 3.1. This proof uses some of the results developed in the proof of Theorem 1 of Chapter 2. Let us first define some notation. Let $\boldsymbol{H} = (\boldsymbol{\eta}^{(1)}, \dots, \boldsymbol{\eta}^{(p)})$, $\hat{\boldsymbol{H}} = (\hat{\boldsymbol{\eta}}^{(1)}, \dots, \hat{\boldsymbol{\eta}}^{(p)})$. Furthermore, Let $\mathcal{P}(\boldsymbol{A}) = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$ denote the projection on the space spanned by $\boldsymbol{A}$ and let $\mathcal{M}(\boldsymbol{A}) = I - \mathcal{P}(\boldsymbol{A})$ denote the corresponding residual-maker. For any matrix $\boldsymbol{A}$, let the norm $\|\cdot\|_p$ represent the induced l_p-matrix norm $\|\boldsymbol{A}\|_p = \sup_{x \neq 0} \|\boldsymbol{A}x\|_p / \|x\|_p$.

Consider the true, unobserved, DGP equation in (3.19) restated here:

$$\boldsymbol{y} = \boldsymbol{X}_{\underline{GC}}\boldsymbol{\beta}_{\underline{GC}} + \boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC} + \boldsymbol{\epsilon}. \tag{3.54}$$

As done in (3.24) in Section 3.2, we rewrite equation (3.54) by augmenting d extra lags of the Granger causing and Granger caused variables to get the $T \times N_\varphi(p+d)$ matrix $\boldsymbol{X}^*_{\underline{GC}}$ where recall $N_\varphi = N_J + 1$. As this is the true DGP, the corresponding $N_\varphi(p+d) \times 1$ augmented coefficient vector $\boldsymbol{\beta}^*_{\underline{GC}}$ will correspond to the $N_\varphi p \times 1$ vector $\boldsymbol{\beta}_{\underline{GC}}$ with zeroes in place of the coefficient of the extra d lags. Thus, by using the

$N_\varphi(p+d) \times N_\varphi(p+d)$ transformation matrices $\boldsymbol{P}_d$ as in Section 3.2 we obtain

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}^*_{\underline{GC}}\boldsymbol{\beta}^*_{\underline{GC}} + \boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC} + \boldsymbol{u} \\ &= \boldsymbol{X}^*_{\underline{GC}}\boldsymbol{P}_d\boldsymbol{P}_d^{-1}\boldsymbol{\beta}^*_{\underline{GC}} + \boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC} + \boldsymbol{u} \\ &= \boldsymbol{W}^*_d\boldsymbol{\phi}^* + \boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC} + \boldsymbol{u}, \end{aligned} \tag{3.55}$$

where now $\boldsymbol{W}^*_d := \boldsymbol{X}^*_{\underline{GC}}\boldsymbol{P}_d, \boldsymbol{\phi} := \boldsymbol{P}_d^{-1}\boldsymbol{\beta}^*_{\underline{GC}}$. It then follows that our PDS estimator in (3.24) is equivalently recast in terms of $\boldsymbol{W}^*_d$ and $\boldsymbol{\phi}^*$ thus to obtain

$$\begin{aligned} &\hat{\boldsymbol{\phi}}^* = \left(\boldsymbol{W}^{*\prime}_d\mathcal{M}(\boldsymbol{X}_{\hat{S}})\boldsymbol{W}^*_d\right)^{-1}\left(\boldsymbol{W}^{*\prime}_d\mathcal{M}(\boldsymbol{X}_{\hat{S}})\left[\boldsymbol{W}^*_d\boldsymbol{\phi}^* + \boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC} + \boldsymbol{u}\right]\right), \\ &\boldsymbol{D}_T(\hat{\boldsymbol{\phi}}^* - \boldsymbol{\phi}^*) = \\ &= \underbrace{\left(\boldsymbol{D}_T^{-1}(\boldsymbol{W}^{*\prime}_d\mathcal{M}(\boldsymbol{X}_{\hat{S}})\boldsymbol{W}^*_d)\boldsymbol{D}_T^{-1}\right)^{-1}}_{A}\underbrace{\boldsymbol{D}_T^{-1}\left(\boldsymbol{W}^{*\prime}_d\mathcal{M}(\boldsymbol{X}_{\hat{S}})\left[\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC} + \boldsymbol{u}\right]\right)}_{B}. \end{aligned} \tag{3.56}$$

Now let $\{\boldsymbol{S}_0\}$ be the active set i.e., the set of the truly non-zero variables in $\boldsymbol{X}$. In other words, $\boldsymbol{X}_{\boldsymbol{S}_0} \equiv \boldsymbol{X}_{-GC}$ in the true DGP equations (3.54), (3.55). We assume the cardinality of $\{\boldsymbol{S}_0\}$ to be fixed and not growing with the sample size. We are now going to show in turn that

$$A - \left(\boldsymbol{D}_T^{-1}(\boldsymbol{W}^{*\prime}_d\mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\boldsymbol{W}^*_d)\boldsymbol{D}_T^{-1}\right)^{-1} = o_p(1) \tag{3.57}$$

$$B - \boldsymbol{D}_T^{-1}\left(\boldsymbol{W}^{*\prime}_d\mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\left[\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC} + \boldsymbol{u}\right]\right) = o_p(1). \tag{3.58}$$

Before turning to the proofs, recall $\boldsymbol{X}^{(j)}_{GC}$ is the $T \times 1$ vector for the jth lag of only the Granger causing variable and similarly $\boldsymbol{X}_{-GCj}$ is the the $T \times (N_\varphi p - 1)$ submatrix of $\boldsymbol{X}_{\underline{GC}}$ where the jth lag of the Granger causing has been taken out. Consider taking the following linear projection of $\boldsymbol{X}^{(j)}_{GC}$ onto the space spanned by the columns of $\boldsymbol{X}_{-GCj}$ and $\boldsymbol{X}_{-GC}$:

$$\boldsymbol{X}^{(j)}_{GC} = \boldsymbol{X}_{-GCj}\boldsymbol{\eta}_{-GCj} + \boldsymbol{X}_{-GC}\boldsymbol{\eta}^{(j)}_{-GC} + \boldsymbol{X}^{(j)}_{GC}\boldsymbol{0} + \boldsymbol{e}^{(j)}, \tag{3.59}$$

for $j = 1, \ldots, (N_\varphi - 1)p$, where $\boldsymbol{\eta}_{-GCj}$ is $(N_\varphi p - 1) \times 1$, $\boldsymbol{\eta}_{-GC}$ is $p(K - N_\varphi) \times 1$ and $\boldsymbol{0}$ is 1×1. Now let the $T \times (N_\varphi - 1)p$ matrix $\boldsymbol{X}_{GC} =$

$(\boldsymbol{X}_{GC}^{(1)},\dots,\boldsymbol{X}_{GC}^{(p)})$, and the $T\times pK$ matrix $\boldsymbol{X}=\left[\boldsymbol{X}_{-GCj},\boldsymbol{X}_{GC}^{(j)},\boldsymbol{X}_{-GC}\right]$, define the $pK\times 1$ vector $\underline{\boldsymbol{\eta}}^{(j)}=\left(\boldsymbol{\eta}_{-GCj}{}',\mathbf{0}',\boldsymbol{\eta}'_{-GC}\right)'$ and its $pK\times(N_\varphi-1)p$ stacked form $\boldsymbol{H}=\left(\underline{\boldsymbol{\eta}}^{(1)},\dots,\underline{\boldsymbol{\eta}}^{(p(N_\varphi-1))}\right)$ and finally the $T\times(N_\varphi-1)p$ stacked matrix $\boldsymbol{e}=\left(\boldsymbol{e}^{(1)}\dots,\boldsymbol{e}^{(p(N_\varphi-1))}\right)$ such that we have the stacked projections

$$\boldsymbol{X}_{GC}=\boldsymbol{X}\boldsymbol{H}+\boldsymbol{e}. \tag{3.60}$$

Using this projection allow us to obtain an expression for the $T\times N_\varphi(p+d)$ matrix $\boldsymbol{X}_{\underline{GC}}^{*}$, namely

$$\begin{aligned}\boldsymbol{X}_{\underline{GC}}^{*}&=[\boldsymbol{X},\boldsymbol{X}_{GC}^{+}]\begin{bmatrix}\boldsymbol{H} & \boldsymbol{I}_{pN_I} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}_{N_\varphi d}\end{bmatrix}+[\boldsymbol{e},\mathbf{0},\mathbf{0}]\\ &=\left[\boldsymbol{X}\boldsymbol{H}+\boldsymbol{e},\boldsymbol{X}_{pN_I,GC^{(+)}}\right],\end{aligned} \tag{3.61}$$

where we use $\boldsymbol{X}_{GC}^{+}$ to indicate the $T\times N_\varphi d$ submatrix of $\boldsymbol{X}_{\underline{GC}}^{*}$ only containing the extra d lags of the Granger causing and Granger caused variables; $\boldsymbol{I}_{N_I}$ is a $pK\times p$ matrix made of $\mathbf{0}$ and $\boldsymbol{I}$ which extracts from $\boldsymbol{X}$ the columns corresponding to the p lags of the Granger caused variable(s)[18] i.e., $\boldsymbol{X}\boldsymbol{I}_{pN_I}=\boldsymbol{X}_{pN_I}$; $\boldsymbol{X}_{pN_I,GC^{(+)}}$ indicates the joint $T\times(p+N_\varphi d)$ containing both $\boldsymbol{X}\boldsymbol{I}_{pN_I}$ and $\boldsymbol{X}_{GC}^{+}\boldsymbol{I}_{N_\varphi d}$. Also, consider the interaction among $\boldsymbol{X}_{\underline{GC}}^{*}$ and the $N_\varphi(p+d)\times N_\varphi(p+d)$ matrix $\boldsymbol{P}_d$:

$$\begin{aligned}\boldsymbol{X}_{\underline{GC}}^{*}\boldsymbol{P}_d&=\left[\boldsymbol{X}\boldsymbol{H}+\boldsymbol{e},\boldsymbol{X}_{pN_I,GC^{(+)}}\right]\begin{bmatrix}\boldsymbol{P}_d^{(1)}\\ \boldsymbol{P}_d^{(2)}\end{bmatrix}\\ &=\left(\boldsymbol{X}\boldsymbol{H}+\boldsymbol{e}\right)\boldsymbol{P}_d^{(1)}+\boldsymbol{X}_{pN_I,GC^{(+)}}\boldsymbol{P}_d^{(2)},\end{aligned} \tag{3.62}$$

for $\boldsymbol{P}_d^{(1)}$ a $p\times N_\varphi(p+d)$ matrix and $\boldsymbol{P}_d^{(2)}$ a $(p+N_\varphi d)\times N_\varphi(p+d)$. When occasionally needed in the proof we will refer to $\boldsymbol{P}_{d,1}^{(1)},\boldsymbol{P}_{d,1}^{(2)}$ as the $1\times N_\varphi(p+d)$ subparts of these matrices. Likewise for $\boldsymbol{D}_{T,pK}^{-1}$, $\boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1}$.

[18] We are using this more general notation pN_I even though we are working under the assumption that $N_I=1$ just to highlight how the procedure can also be shown for a block of Granger caused variables.

As the interaction between $\boldsymbol{D}_T^{-1}$ and $\boldsymbol{P}_d$ occurs often in the proof, we report here an example of their 3×3 matrix product as a reference:

$$\boldsymbol{P}_1\boldsymbol{D}_T^{-1} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{T} & 0 & 0 \\ 0 & 1/T & 0 \\ 0 & 0 & 1/T^2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{T} & 0 & 1/T^2 \\ 0 & 1/T & 0 \\ 0 & 0 & 1/T^2 \end{bmatrix},$$

$$\boldsymbol{P}_2\boldsymbol{D}_T^{-1} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{T} & 0 & 0 \\ 0 & 1/T & 0 \\ 0 & 0 & 1/T^2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{T} & 0 & 2/T^2 \\ 0 & 1/T & 0 \\ 0 & 0 & 1/T^2 \end{bmatrix}.$$

We now prove (3.57). By rearranging, we have

$$\boldsymbol{D}_T^{-1}\boldsymbol{W}_d^{*\prime}\left(\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}}) - \mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\right)\boldsymbol{W}_d^{*}\boldsymbol{D}_T^{-1}.$$

Using that $\boldsymbol{W}_d^{*} = \boldsymbol{X}_{\underline{GC}}^{*}\boldsymbol{P}_d$ and using the expression in (3.62) for $\boldsymbol{X}_{\underline{GC}}^{*}$ we obtain the following expression:

$$\boldsymbol{D}_T^{-1}\left(\left(\boldsymbol{X}\boldsymbol{H} + \boldsymbol{e}\right)\boldsymbol{P}_d^{(1)} + \boldsymbol{X}_{pN_I,GC^{(+)}}\boldsymbol{P}_d^{(2)}\right)^{\prime}$$
$$\left(\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}}) - \mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\right)\left(\left(\boldsymbol{X}\boldsymbol{H} + \boldsymbol{e}\right)\boldsymbol{P}_d^{(1)} + \boldsymbol{X}_{pN_I,GC^{(+)}}\boldsymbol{P}_d^{(2)}\right)\boldsymbol{D}_T^{-1}.$$

By multiplying out the expression we obtain three terms:

$$\underbrace{\boldsymbol{D}_T^{-1}\left(\left(\boldsymbol{X}\boldsymbol{H} + \boldsymbol{e}\right)\boldsymbol{P}_d^{(1)}\right)^{\prime}\left(\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}}) - \mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\right)\left(\left(\boldsymbol{X}\boldsymbol{H} + \boldsymbol{e}\right)\boldsymbol{P}_d^{(1)}\right)\boldsymbol{D}_T^{-1}}_{A_1} +$$
$$+\underbrace{\boldsymbol{D}_T^{-1}\left(\boldsymbol{X}_{pN_I,GC^{(+)}}\boldsymbol{P}_d^{(2)}\right)^{\prime}\left(\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}}) - \mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\right)\left(\boldsymbol{X}_{pN_I,GC^{(+)}}\boldsymbol{P}_d^{(2)}\right)\boldsymbol{D}_T^{-1}}_{A_2} +$$
$$+\,2\underbrace{\boldsymbol{D}_T^{-1}\left(\left(\boldsymbol{X}\boldsymbol{H} + \boldsymbol{e}\right)\boldsymbol{P}_d^{(1)}\right)^{\prime}\left(\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}}) - \mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\right)\left(\boldsymbol{X}_{pN_I,GC^{(+)}}\boldsymbol{P}_d^{(2)}\right)\boldsymbol{D}_T^{-1}}_{A_3}.$$

Let us deal with A_1 first. Multiplying out the terms and using triangle inequality within an ℓ_2-norm we get:

$$\begin{aligned}\|A_1\|_2 \leq & \underbrace{\left\| \boldsymbol{D}_T^{-1}(\boldsymbol{XHP}_d^{(1)})' \left(\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}}) - \mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\right) \boldsymbol{XHP}_d^{(1)} \boldsymbol{D}_T^{-1} \right\|_2}_{A_{1,1}} + \\ & + \underbrace{\left\| \boldsymbol{D}_T^{-1}(\boldsymbol{eP}_d^{(1)})' \left(\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}}) - \mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\right) (\boldsymbol{eP}_d^{(1)}) \boldsymbol{D}_T^{-1} \right\|_2}_{A_{1,2}} + \\ & + \underbrace{2\left\| \boldsymbol{D}_T^{-1}(\boldsymbol{XHP}_d^{(1)})' \left(\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}}) - \mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\right) (\boldsymbol{eP}_d^{(1)}) \boldsymbol{D}_T^{-1} \right\|_2}_{A_{1,3}}.\end{aligned}$$

For $A_{1,1}$, observe how $\boldsymbol{X}$ is only active on the active set $\{\boldsymbol{S}_0\}$, such that $\mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\boldsymbol{XH} = \boldsymbol{X}_{\boldsymbol{S}_0}\boldsymbol{H}_{\boldsymbol{S}_0} - \mathcal{P}(\boldsymbol{X}_{\boldsymbol{S}_0})\boldsymbol{X}_{\boldsymbol{S}_0}\boldsymbol{H}_{\boldsymbol{S}_0} = 0$. Hence, $A_{1,1}$ reduces to

$$\begin{aligned} A_{1,1} = & \left\| \mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}})\boldsymbol{XHP}_d^{(1)} \boldsymbol{D}_T^{-1} \right\|_2^2 \overset{(I)}{\leq} \left\| \boldsymbol{P}_d^{(1)} \boldsymbol{D}_T^{-1} \right\|_2^2 \left\| \mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}})\boldsymbol{XH} \right\|_2^2 \\ \overset{(II)}{\leq} & \left\| \boldsymbol{P}_d^{(1)} \boldsymbol{D}_T^{-1} \right\|_2^2 \sum_{j=1}^{(N_\varphi - 1)p} \left\| \mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}})\boldsymbol{X}\underline{\boldsymbol{\eta}}^{(j)} \right\|_2^2 \\ \overset{(III)}{\leq} & \left\| \boldsymbol{P}_d^{(1)} \boldsymbol{D}_T^{-1} \right\|_2^2 \sum_{j=1}^{(N_\varphi - 1)p} \left\| \mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}_j})\boldsymbol{X}\underline{\boldsymbol{\eta}}^{(j)} \right\|_2^2, \end{aligned}$$

where (I) follows from submultiplicativity of the induced norm, (II) follows by applying triangle inequality and (III) from observing that $\{\hat{\boldsymbol{S}}_j\} \subseteq \{\hat{\boldsymbol{S}}\}$ for all $j = 1, \dots, p$. Furthermore,

$$\begin{aligned} & \left\| \boldsymbol{P}_d^{(1)} \boldsymbol{D}_T^{-1} \right\|_2^2 \left\| \mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}_j})\boldsymbol{X}\underline{\boldsymbol{\eta}}^{(j)} \right\|_2^2 = \\ & = \left\| \boldsymbol{P}_d^{(1)} \boldsymbol{D}_T^{-1} \right\|_2^2 \min_{\underline{\boldsymbol{\eta}} : \underline{\boldsymbol{\eta}}_m = 0, m \notin \hat{\boldsymbol{S}}_j} \left\| \boldsymbol{X}\underline{\boldsymbol{\eta}}^{(j)} - \boldsymbol{X}_{\hat{\boldsymbol{S}}_j}\underline{\boldsymbol{\eta}} \right\|_2^2 \\ & \leq \left\| \boldsymbol{P}_d^{(1)} \boldsymbol{D}_T^{-1} \right\|_2^2 \sqrt{T} \left\| \boldsymbol{X}(\underline{\boldsymbol{\eta}}^{(j)} - \hat{\underline{\boldsymbol{\eta}}}^{(j)}) \right\|_2^2 \leq \delta_T^2, \end{aligned}$$

as the constraint in the minimization is satisfied given $\hat{\boldsymbol{S}}_j = \{m : \underline{\hat{\boldsymbol{\eta}}}_i^{(j)} = 0\}$ and where the last inequality follows for the first part given the maximum eigenvalue of the induced matrix norm squared is T^{-1} thus majorated by $T^{-1/2}$; the second part follows with probability $1 - \Delta_T$ from Assumption 4(e).

For $A_{1,2}$, by rewriting the residual makers as the difference between the identity and the corresponding projection matrices, we simplify the expression to

$$\begin{aligned} A_{1,2} &= \left\| \boldsymbol{D}_T^{-1}(\boldsymbol{e}\boldsymbol{P}_d^{(1)})' \left(\mathcal{P}(\boldsymbol{X}_{\boldsymbol{S}_0}) - \mathcal{P}(\boldsymbol{X}_{\hat{\boldsymbol{S}}})\right)(\boldsymbol{e}\boldsymbol{P}_d^{(1)})\boldsymbol{D}_T^{-1} \right\|_2 \\ &\leq \underbrace{\left\| \boldsymbol{D}_T^{-1}(\boldsymbol{e}\boldsymbol{P}_d^{(1)})'\mathcal{P}(\boldsymbol{X}_{\boldsymbol{S}_0})(\boldsymbol{e}\boldsymbol{P}_d^{(1)})\boldsymbol{D}_T^{-1} \right\|_2}_{A_{1,2,1}} + \\ &+ \underbrace{\left\| \boldsymbol{D}_T^{-1}(\boldsymbol{e}\boldsymbol{P}_d^{(1)})'\mathcal{P}(\boldsymbol{X}_{\hat{\boldsymbol{S}}})(\boldsymbol{e}\boldsymbol{P}_d^{(1)})\boldsymbol{D}_T^{-1} \right\|_2}_{A_{1,2,2}}, \end{aligned}$$

where

$$\begin{aligned} A_{1,2,1} &\overset{(I)}{\leq} \left\| \boldsymbol{D}_T^{-1}(\boldsymbol{e}\boldsymbol{P}_d^{(1)})'\boldsymbol{X}_{\boldsymbol{S}_0}\boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1}(\boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1}\boldsymbol{X}_{\boldsymbol{S}_0}'\boldsymbol{X}_{\boldsymbol{S}_0}\boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1})^{-1}\boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1}\boldsymbol{X}_{\boldsymbol{S}_0}'(\boldsymbol{e}\boldsymbol{P}_d^{(1)})\boldsymbol{D}_T^{-1} \right\|_1 \\ &\overset{(II)}{\leq} \left\| \boldsymbol{D}_{T,pK}^{-1}\boldsymbol{X}'(\boldsymbol{e}\boldsymbol{P}_d^{(1)})\boldsymbol{D}_T^{-1} \right\|_\infty \\ &\sum_{j=1}^{p(N_\varphi-1)} \sqrt{\bar{s}_T}\left\| (\boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1}\boldsymbol{X}_{\boldsymbol{S}_0}'\boldsymbol{X}_{\boldsymbol{S}_0}\boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1})^{-1}\boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1}\boldsymbol{X}_{\boldsymbol{S}_0}'(\boldsymbol{e}^{(j)}\boldsymbol{P}_{d,1}^{(1)})\boldsymbol{D}_T^{-1} \right\|_2 \\ &\overset{(III)}{\leq} \left\| \boldsymbol{D}_{T,pK}^{-1}\boldsymbol{X}'(\boldsymbol{e}\boldsymbol{P}_d^{(1)})\boldsymbol{D}_T^{-1} \right\|_\infty \|(\boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1}\boldsymbol{X}_{\boldsymbol{S}_0}'\boldsymbol{X}_{\boldsymbol{S}_0}\boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1})^{-1}\|_2 \\ &\sum_{j=1}^{p(N_\varphi-1)} \bar{s}_T\left\| \boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1}\boldsymbol{X}_{\boldsymbol{S}_0}'(\boldsymbol{e}^{(j)}\boldsymbol{P}_{d,1}^{(1)})\boldsymbol{D}_T^{-1} \right\|_\infty \\ &\overset{(IV)}{\leq} \|\boldsymbol{D}_{T,pK}^{-1}\boldsymbol{X}'\boldsymbol{e}\|_\infty \left\| \boldsymbol{P}_d^{(1)}\boldsymbol{D}_T^{-1} \right\|_\infty \|(\boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1}\boldsymbol{X}_{\boldsymbol{S}_0}'\boldsymbol{X}_{\boldsymbol{S}_0}\boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1})^{-1}\|_2 \\ &\sum_{j=1}^{p(N_\varphi-1)} \bar{s}_T\left\| \boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1}\boldsymbol{X}_{\boldsymbol{S}_0}'(\boldsymbol{e}^{(j)}\boldsymbol{P}_{d,1}^{(1)})\boldsymbol{D}_T^{-1} \right\|_\infty \end{aligned}$$

$$
\begin{aligned}
&\overset{(V)}{\leq} p(N_\varphi - 1)\bar{s}_T \boldsymbol{\kappa}_{T,\min}^{-1} \left\| \boldsymbol{P}_d^{(1)} \boldsymbol{D}_T^{-1} \right\|_\infty^2 \left\| \boldsymbol{D}_{T,pK}^{-1} \boldsymbol{X}' \boldsymbol{e} \right\|_\infty \left\| \boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1} \boldsymbol{X}_{\boldsymbol{S}_0}' \boldsymbol{e} \right\|_\infty \\
&\overset{(VI)}{\leq} p(N_\varphi - 1)\bar{\gamma}_T^2 \bar{s}_T \boldsymbol{\kappa}_{T,\min}^{-1} T^{-1} \leq \delta_T^2,
\end{aligned}
$$

where (I) follows by bounding ℓ_2 with ℓ_1 norm; (II) follows from the dual norm inequality: for any $m \times n$ matrix $\boldsymbol{A}$ with i-th row denoted as $a_{i\cdot}$, and $n \times 1$ vector $\boldsymbol{x}$, we have that $\|\boldsymbol{A}\boldsymbol{x}\|_1 = \sum_{i=1}^m |a_{i\cdot}x| \leq \|\boldsymbol{x}\|_\infty \sum_{i=1}^m \|a_i\|_1$. The first term majorate $\boldsymbol{X}_{\boldsymbol{S}_0} \boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1}$ into $\boldsymbol{D}_{T,pK}^{-1} \boldsymbol{X}'$. Furthermore, we use triangle inequality on the $\boldsymbol{e}$'s and bound l_1 by l_2 norms which let the $\sqrt{\bar{s}_T}$ appear; (III) follows from Cauchy-Schwarz and bounding the second term in ℓ_2 norm by ℓ_∞ norm; (IV) follows by submultiplicativity of the induced norm. In (V) the central term in ℓ_2 norm at previous step is bounded according to Assumption 4,(g). Union bound is applied on the last term yielding the $p(N_\varphi - 1)$ term while similar terms in $\boldsymbol{P}_d^{(1)} \boldsymbol{D}_T^{-1}$ are aggregated. Finally in (VI) the empirical process bound in Assumption 4,(c) is used on the last two terms yielding the $\bar{\gamma}_T^2$ and the T^{-1} comes from the absolute maximal row sum of the remaining squared matrix norm. In the same way as for $A_{1,1,1}$, we get that $A_{1,2,2} \leq \delta_T^2$.

Finally for $A_{1,3}$. By the same argument as in $A_{1,1}$ we get that $A_{1,3}$ reduces to

$$
A_{1,3} = 2 \left\| \boldsymbol{D}_T^{-1} (\boldsymbol{X}\boldsymbol{H}\boldsymbol{P}_d^{(1)})' \mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}}) (\boldsymbol{e}\boldsymbol{P}_d^{(1)}) \boldsymbol{D}_T^{-1} \right\|_2.
$$

Let us define the noiseless least squares estimator

$$
\underline{\tilde{\boldsymbol{\eta}}}_{\hat{S}}^{(j)} = \underset{\underline{\boldsymbol{\eta}}:\underline{\boldsymbol{\eta}}_m = 0, m \notin \hat{\boldsymbol{S}}}{\arg\min} \left\| \boldsymbol{X}\underline{\boldsymbol{\eta}}^{(j)} - \boldsymbol{X}\underline{\boldsymbol{\eta}} \right\|_2^2, \qquad j = 0, 1, \ldots, p, \qquad (3.63)
$$

and let $\tilde{\boldsymbol{H}}_{\hat{\boldsymbol{S}}} = \left(\underline{\tilde{\boldsymbol{\eta}}}_{\hat{\boldsymbol{S}}}^{(j)}, \ldots, \underline{\tilde{\boldsymbol{\eta}}}_{\hat{\boldsymbol{S}}}^{(j)} \right)$, such that $\boldsymbol{D}_T^{-1} \left(\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}}) \boldsymbol{X}\boldsymbol{H}\boldsymbol{P}_d^{(1)} \right)' = \boldsymbol{D}_T^{-1} \left(\boldsymbol{X} \left(\boldsymbol{H} - \tilde{\boldsymbol{H}}_{\hat{\boldsymbol{S}}} \right) \boldsymbol{P}_d^{(1)} \right)'$. Then, with probability $1 - \Delta_T$,

$$
A_{1,3} \overset{(I)}{\leq} \left\| \boldsymbol{D}_T^{-1} \left(\boldsymbol{X} \left(\tilde{\boldsymbol{H}}_{\hat{\boldsymbol{S}}} - \boldsymbol{H} \right)' \boldsymbol{P}_d^{(1)} \right)' \boldsymbol{e}\boldsymbol{P}_d^{(1)} \boldsymbol{D}_T^{-1} \right\|_1
$$

$$
\begin{aligned}
&\overset{(II)}{\leq} \sum_{j=1}^{p(N_\varphi-1)} \left\| \boldsymbol{D}_T^{-1} \left(\left(\underline{\tilde{\boldsymbol{\eta}}}_{\hat{\boldsymbol{S}}}^{(j)} - \underline{\boldsymbol{\eta}}^{(j)} \right) \boldsymbol{P}_{d,1}^{(1)} \right)' \right\|_1 \left\| \boldsymbol{D}_{T,pK} \boldsymbol{D}_{T,pK}^{-1} \boldsymbol{X}' \boldsymbol{e} \boldsymbol{P}_d^{(1)} \boldsymbol{D}_T^{-1} \right\|_\infty \\
&\overset{(III)}{\leq} \sum_{j=1}^{p(N_\varphi-1)} \left\| \boldsymbol{D}_T^{-1} \left(\left(\underline{\tilde{\boldsymbol{\eta}}}_{\hat{\boldsymbol{S}}}^{(j)} - \underline{\boldsymbol{\eta}}^{(j)} \right) \boldsymbol{P}_{d,1}^{(1)} \right)' \right\|_1 \left\| \boldsymbol{D}_{T,pK} \right\|_\infty \left\| \boldsymbol{D}_{T,pK}^{-1} \boldsymbol{X}' \boldsymbol{e} \right\|_\infty \left\| \boldsymbol{P}_d^{(1)} \boldsymbol{D}_T^{-1} \right\|_\infty \\
&\overset{(IV)}{\leq} T^{3/2} \bar{\gamma}_T \sum_{j=1}^{p(N_\varphi-1)} \left\| \boldsymbol{D}_T^{-1} \left(\left(\underline{\tilde{\boldsymbol{\eta}}}_{\hat{\boldsymbol{S}}}^{(j)} - \underline{\boldsymbol{\eta}}^{(j)} \right) \boldsymbol{P}_{d,1}^{(1)} \right)' \right\|_1 \\
&\overset{(V)}{\leq} T^{3/2} \frac{\sqrt{\bar{s}_T} \bar{\gamma}_T}{\boldsymbol{\kappa}_{T,\min}} \sum_{j=1}^{p(N_\varphi-1)} \left\| \boldsymbol{D}_T^{-1} \left(\boldsymbol{D}_{T,pK}^{-1} \boldsymbol{X} \left(\underline{\tilde{\boldsymbol{\eta}}}_{\hat{\boldsymbol{S}}}^{(j)} - \underline{\boldsymbol{\eta}}^{(j)} \right) \boldsymbol{P}_{d,1}^{(1)} \right)' \right\|_2 \\
&\overset{(VI)}{\leq} T^{3/2} \frac{\sqrt{\bar{s}_T} \bar{\gamma}_T}{\boldsymbol{\kappa}_{T,\min}} \sum_{j=1}^{p(N_\varphi-1)} \left\| \boldsymbol{D}_T^{-1} \left(\boldsymbol{D}_{T,pK}^{-1} \boldsymbol{X} \left(\underline{\hat{\boldsymbol{\eta}}}^{(j)} - \underline{\boldsymbol{\eta}}^{(j)} \right) \boldsymbol{P}_{d,1}^{(1)} \right)' \right\|_2 \\
&\overset{(VII)}{\leq} T^{3/2} \frac{\sqrt{\bar{s}_T} \bar{\gamma}_T}{\boldsymbol{\kappa}_{T,\min}} \sum_{j=1}^{p(N_\varphi-1)} \left\| \boldsymbol{D}_{T,pK}^{-1} \boldsymbol{X} \left(\underline{\hat{\boldsymbol{\eta}}}^{(j)} - \underline{\boldsymbol{\eta}}^{(j)} \right) \right\|_2 \left\| \boldsymbol{D}_T^{-1} \boldsymbol{P}_{d,1}^{(1)\prime} \right\|_2 \\
&\overset{(VIII)}{\leq} \sqrt{T} \delta_T \sqrt{\bar{s}_T} \bar{\gamma}_T \boldsymbol{\kappa}_{T,\min}^{-1} \leq \delta_T^2.
\end{aligned}
$$

Inequality (I) bounds ℓ_2 with ℓ_1 norm. Inequality (II) follows from the dual norm inequality and by triangle inequality on $\underline{\tilde{\boldsymbol{\eta}}}^{(j)}$'s. Also in the second term $\boldsymbol{D}_{T,pK}\boldsymbol{D}_{T,pK}^{-1}$ is added. (III) follows from submultiplicativity on the second term. (IV) follows from the empirical process bound in Assumption 4(c) for the second term in ℓ_∞ norm in the previous step. The factor $T^{3/2}$ comes from the maximum column sum of the first term in ℓ_∞ norm i.e., T^2 times the maximum column sum of the last term in ℓ_∞ norm i.e., $T^{-1/2}$. Step (V) follows from combining Assumption 4(f) and 1(g), namely sparsity and the restricted eigenvalue condition. (VI) follows from the definition of $\underline{\tilde{\boldsymbol{\eta}}}_{\hat{S}}$ as minimizer of the sum of squares; (VII) follows from the submultiplicativity of the induced norm and finally (VIII) follows from the consistency in Assumption 4(e). Therefore, it follows from combining the previous results that $\|A_1\|_2 = o_p(1)$.

We now prove $A_2 = o_p(1)$. To do so we follow a similar strategy as for $A_{1,2}$. By rewriting the residual makers as the difference between the identity and the corresponding projection matrices, the expression can be simplified to

$$
\begin{aligned}
A_2 &= \left\| \boldsymbol{D}_T^{-1} (\boldsymbol{X}_{pN_I, GC^{(+)}} \boldsymbol{P}_d^{(2)})' \left(\mathcal{P}(\boldsymbol{X}_{\boldsymbol{S}_0}) - \mathcal{P}(\boldsymbol{X}_{\hat{\boldsymbol{S}}}) \right) (\boldsymbol{X}_{pN_I, GC^{(+)}} \boldsymbol{P}_d^{(2)}) \boldsymbol{D}_T^{-1} \right\|_2 \\
&\leq \underbrace{\left\| \boldsymbol{D}_T^{-1} (\boldsymbol{X}_{pN_I, GC^{(+)}} \boldsymbol{P}_d^{(2)})' \mathcal{P}(\boldsymbol{X}_{\boldsymbol{S}_0}) (\boldsymbol{X}_{pN_I, GC^{(+)}} \boldsymbol{P}_d^{(2)}) \boldsymbol{D}_T^{-1} \right\|_2}_{A_{2,1}} + \\
&+ \underbrace{\left\| \boldsymbol{D}_T^{-1} (\boldsymbol{X}_{pN_I, GC^{(+)}} \boldsymbol{P}_d^{(2)})' \mathcal{P}(\boldsymbol{X}_{\hat{\boldsymbol{S}}}) (\boldsymbol{X}_{pN_I, GC^{(+)}} \boldsymbol{P}_d^{(2)}) \boldsymbol{D}_T^{-1} \right\|_2}_{A_{2,2}}.
\end{aligned}
$$

Let us assume that an empirical process bound, similar to Assumption 4, (c) holds such that

$$
\left\| \boldsymbol{D}_{T,pK}^{-1} \boldsymbol{X}' \boldsymbol{X}_{pN_I, GC^{(+)}} \boldsymbol{D}_{T,(p+N_\varphi d))}^{-1} \right\|_\infty \leq 3\bar{\bar{\gamma}}_T. \tag{3.64}
$$

Then, similarly to $A_{1,2,1}$ we have

$$
\begin{aligned}
A_{2,1} \overset{(I)}{\leq} & \Big\| \boldsymbol{D}_T^{-1} (\boldsymbol{X}_{pN_I, GC^{(+)}} \boldsymbol{P}_d^{(2)})' \boldsymbol{X}_{\boldsymbol{S}_0} \boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1} \\
& (\boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1} \boldsymbol{X}_{\boldsymbol{S}_0}' \boldsymbol{X}_{\boldsymbol{S}_0} \boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1})^{-1} \boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1} \boldsymbol{X}_{\boldsymbol{S}_0}' (\boldsymbol{X}_{pN_I, GC^{(+)}} \boldsymbol{P}_d^{(2)}) \Big\|_1 \\
\overset{(II)}{\leq} & \left\| \boldsymbol{D}_{T,pK}^{-1} \boldsymbol{X}' (\boldsymbol{X}_{pN_I, GC^{(+)}} \boldsymbol{P}_d^{(2)}) \boldsymbol{D}_T^{-1} \right\|_\infty \\
& \sum_{j=1}^{p+N_\varphi d} \sqrt{\bar{s}_T} \left\| (\boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1} \boldsymbol{X}_{\boldsymbol{S}_0}' \boldsymbol{X}_{\boldsymbol{S}_0} \boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1})^{-1} \boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1} \boldsymbol{X}_{\boldsymbol{S}_0}' (\boldsymbol{X}_{pN_I, GC^{(+)}}^{(j)} \boldsymbol{P}_{d,1}^{(2)}) \boldsymbol{D}_T^{-1} \right\|_2 \\
\overset{(III)}{\leq} & \left\| \boldsymbol{D}_{T,pK}^{-1} \boldsymbol{X}' (\boldsymbol{X}_{pN_I, GC^{(+)}} \boldsymbol{P}_d^{(2)}) \boldsymbol{D}_T^{-1} \right\|_\infty \left\| (\boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1} \boldsymbol{X}_{\boldsymbol{S}_0}' \boldsymbol{X}_{\boldsymbol{S}_0} \boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1})^{-1} \right\|_2 \\
& \sum_{j=1}^{p+N_\varphi d} \bar{s}_T \left\| \boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1} \boldsymbol{X}_{\boldsymbol{S}_0}' (\boldsymbol{X}_{pN_I, GC^{(+)}}^{(j)} \boldsymbol{P}_{d,1}^{(2)}) \boldsymbol{D}_T^{-1} \right\|_\infty \\
\overset{(IV)}{\leq} & \left\| \boldsymbol{D}_{T,pK}^{-1} \boldsymbol{X}' \boldsymbol{X}_{pN_I, GC^{(+)}} \right\|_\infty \left\| \boldsymbol{P}_d^{(1)} \boldsymbol{D}_T^{-1} \right\|_\infty \left\| (\boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1} \boldsymbol{X}_{\boldsymbol{S}_0}' \boldsymbol{X}_{\boldsymbol{S}_0} \boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1})^{-1} \right\|_2 \\
& \sum_{j=1}^{p+N_\varphi d)} \bar{s}_T \left\| \boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1} \boldsymbol{X}_{\boldsymbol{S}_0}' (\boldsymbol{X}_{pN_I, GC^{(+)}}^{(j)} \boldsymbol{P}_{d,1}^{(2)}) \boldsymbol{D}_T^{-1} \right\|_\infty \\
\overset{(V)}{\leq} & (p + N_\varphi d) \bar{s}_T \boldsymbol{\kappa}_{T,\min}^{-1} \left\| \boldsymbol{P}_d^{(1)} \boldsymbol{D}_T^{-1} \right\|_\infty^2 \left\| \boldsymbol{D}_{T,pK}^{-1} \boldsymbol{X}' \boldsymbol{X}_{pN_I, GC^{(+)}} \boldsymbol{D}_{T,(p+N_\varphi d))}^{-1} \right\|_\infty
\end{aligned}
$$

$$\left\|\boldsymbol{D}_{T,\boldsymbol{S}_0}^{-1}\boldsymbol{X}_{\boldsymbol{S}_0}'\boldsymbol{X}_{pN_I,GC^{(+)}}\boldsymbol{D}_{T,(p+N_\varphi d))}^{-1}\right\|_\infty$$
$$\overset{(VI)}{\leq} (p+N_\varphi d)\bar{\bar{\gamma}}_T^2\bar{s}_T\boldsymbol{\kappa}_{T,\min}^{-1}T^3 \leq \delta_T^2.$$

Likewise, it follows $A_{2,2} \leq \delta_T^2$.

Now for A_3.

$$\|A_3\|_2 \leq \underbrace{\left\|\boldsymbol{D}_T^{-1}\left(\boldsymbol{XHP}_d^{(1)}\right)'(\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}}) - \mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0}))\left(\boldsymbol{X}_{pN_I,GC^{(+)}}\boldsymbol{P}_d^{(2)}\right)\boldsymbol{D}_T^{-1}\right\|_2}_{A_{3,1}} +$$
$$+\underbrace{\left\|\boldsymbol{D}_T^{-1}\left(\boldsymbol{eP}_d^{(1)}\right)'(\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}}) - \mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0}))\left(\boldsymbol{X}_{pN_I,GC^{(+)}}\boldsymbol{P}_d^{(2)}\right)\boldsymbol{D}_T^{-1}\right\|_2}_{A_{3,2}},$$

$A_{3,1}$ follows as in $A_{1,3}$. $A_{3,2}$ follows as $A_{1,2,1}$.

Let us now prove (3.58). By re-arranging terms we have

$$\underbrace{\boldsymbol{D}_T^{-1}\boldsymbol{W}_d^{*\prime}\left(\mathcal{M}(\boldsymbol{X}_{\hat{S}}) - \mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\right)\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC}}_{B_1} +$$
$$+\underbrace{\boldsymbol{D}_T^{-1}\boldsymbol{W}_d^{*\prime}\left(\mathcal{M}(\boldsymbol{X}_{\hat{S}}) - \mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\right)\boldsymbol{u}}_{B_2}.$$

We first deal with B_1. Given $\boldsymbol{X}_{-GC}$ is only active on $\{\boldsymbol{S}_0\}$, then $\mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC} = 0$ and B_1 reduces to

$$B_1 = \boldsymbol{D}_T^{-1}\boldsymbol{W}_d^{*\prime}\mathcal{M}(\boldsymbol{X}_{\hat{S}})\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC}.$$

As above for (3.57), using that $\boldsymbol{W}_d^* = \boldsymbol{X}_{GC}^*\boldsymbol{P}_d$ and using the expression in (3.62) for $\boldsymbol{X}_{GC}^*$ we obtain the following expression:

$$\begin{aligned}&\underbrace{\left((\boldsymbol{XH}+\boldsymbol{e})\boldsymbol{P}_d^{(1)}\boldsymbol{D}_T^{-1}\right)'\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}})\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC}}_{B_{1,1}} + \\ &+\underbrace{\left(\boldsymbol{X}_{pN_I,GC^{(+)}}\boldsymbol{P}_d^{(2)}\boldsymbol{D}_T^{-1}\right)'\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}})\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC}}_{B_{1,2}}.\end{aligned} \tag{3.65}$$

By multiplying out the terms in $B_{1,1}$ we have

$$\underbrace{\boldsymbol{D}_T^{-1}(\boldsymbol{X}\boldsymbol{H}\boldsymbol{P}_d^{(1)})'\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}})\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC}}_{B_{1,1,1}}+\underbrace{\boldsymbol{D}_T^{-1}(\boldsymbol{e}\boldsymbol{P}_d^{(1)})'\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}})\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC}}_{B_{1,1,2}}.$$

First for $B_{1,1,1}$. Consider the norm

$$\|B_{1,1,1}\|_2 \leq \underbrace{\sum_{j=1}^{p(N_\varphi-1)} \left\|\boldsymbol{D}_T^{-1}\left(\boldsymbol{X}\underline{\boldsymbol{\eta}}^{(j)}\boldsymbol{P}_{d,1}^{(1)}\right)'\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}})\right\|_2}_{B_{1,1,1,a}} \underbrace{\left\|\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}})\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC}\right\|_2}_{B_{1,1,1,b}}, \tag{3.66}$$

then, $B_{1,1,1,a}$ follows analogously to $A_{1,1}$ such that $\|B_{1,1,a}\| \leq T^{-1/2}\delta_T^2$. For $B_{1,1,1,b}$: note first that we can rewrite $\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC} = \boldsymbol{X}\boldsymbol{\beta}^{**}$ where $\boldsymbol{\beta}^{**}$ contains zeroes in the corresponding places of the GC variables. Then, using the definition of the best linear predictor we have

$$\begin{aligned}\boldsymbol{\beta}^{(0)} &= \left(\bar{\mathbb{E}}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\bar{\mathbb{E}}\boldsymbol{X}'(\boldsymbol{X}_{GC}\boldsymbol{\beta}_{GC}+\boldsymbol{X}\boldsymbol{\beta}^{**}+\boldsymbol{\epsilon}),\\ &= \left(\bar{\mathbb{E}}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\bar{\mathbb{E}}\boldsymbol{X}'\boldsymbol{X}_{GC}\boldsymbol{\beta}_{GC}+\boldsymbol{\beta}^{**} = \boldsymbol{H}\boldsymbol{\beta}_{GC}+\boldsymbol{\beta}^{*},\end{aligned}$$

such that

$$B_{1,1,1,b} \leq \left\|\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}})\boldsymbol{X}\boldsymbol{\beta}^{(0)}\right\|_2 + \left\|\mathcal{M}(\boldsymbol{X}_{\hat{\boldsymbol{S}}})\boldsymbol{X}\boldsymbol{H}\right\|_2\|\boldsymbol{\beta}_{GC}\|_2,$$

where both terms follow similarly and the second can be bouded as

$$\left\|\mathcal{M}\left(\boldsymbol{X}_{\hat{\boldsymbol{S}}}\right)\boldsymbol{X}\boldsymbol{H}\right\|_2\|\boldsymbol{\beta}_{GC}\|_2 \leq C\sum_{j=1}^{p(N_\varphi-1)}\left\|\boldsymbol{X}\left(\boldsymbol{\eta}^{(j)}-\hat{\boldsymbol{\eta}}^{(j)}\right)\right\|_2 \leq Cp\delta_T T^{-1/4}.$$

Now for $B_{1,1,2}$. From the same definition of the best linear predictor as above, it follows that with probability $1-\Delta_T$

$$\begin{aligned}\|B_{1,1,2}\|_1 &\overset{(I)}{\leq} \left\|\boldsymbol{D}_T^{-1}(\boldsymbol{e}\boldsymbol{P}_d^{(1)})'\mathcal{M}(\boldsymbol{X}_{\hat{S}})\boldsymbol{X}\boldsymbol{\beta}^{(0)}\right\|_1 + \left\|\boldsymbol{D}_T^{-1}(\boldsymbol{e}\boldsymbol{P}_d^{(1)})'\mathcal{M}(\boldsymbol{X}_{\hat{S}})\boldsymbol{X}\boldsymbol{H}\boldsymbol{\beta}_{GC}\right\|_1\\ &\overset{(II)}{\leq} \left\|\boldsymbol{D}_T^{-1}(\boldsymbol{e}\boldsymbol{P}_d^{(1)})'\boldsymbol{X}\left(\tilde{\boldsymbol{\beta}}_{\hat{S}}^{(0)}-\boldsymbol{\beta}^{(0)}\right)\right\|_1 + \left\|\boldsymbol{D}_T^{-1}(\boldsymbol{e}\boldsymbol{P}_d^{(1)})'\boldsymbol{X}\left(\tilde{\boldsymbol{H}}_{\hat{S}}-\boldsymbol{H}\right)\boldsymbol{\beta}_{GC}\right\|_1\end{aligned}$$

$$
\begin{aligned}
&\overset{(III)}{\leq} \sum_{j=1}^{p(N_\varphi-1)} \left\| \boldsymbol{D}_T^{-1}(\boldsymbol{e}^{(j)}\boldsymbol{P}_{d,1}^{(1)})'\boldsymbol{X} \right\|_\infty \\
&\left(\left\| \tilde{\boldsymbol{\beta}}_{\hat{S}}^{(0)} - \boldsymbol{\beta}^{(0)} \right\|_1 + \|\boldsymbol{\beta}_{GC}\|_\infty \sum_{j=1}^{p(N_\varphi-1)} \left\| \tilde{\boldsymbol{\eta}}_{\hat{S}}^{(j)} - \boldsymbol{\eta}^{(j)} \right\|_1 \right) \\
&\overset{(IV)}{\leq} T^{-1/2} \frac{\sqrt{\bar{s}_T}\bar{\gamma}_T}{\boldsymbol{\kappa}_{T,\min}} p(N_\varphi - 1)\left(1 + \|\boldsymbol{\beta}_{GC}\|_\infty\right) \sum_{j=0}^{p(N_\varphi-1)} \left\| \boldsymbol{X}\left(\tilde{\boldsymbol{\eta}}_{\hat{S}}^{(j)} - \boldsymbol{\eta}^{(j)} \right) \right\|_2 \\
&\overset{(V)}{\leq} \delta_T T^{-1}\sqrt{\bar{s}_T}\bar{\gamma}_T(C+1)p(N_\varphi-1)(p(N_\varphi-1)+1)\boldsymbol{\kappa}_{T,\min}^{-1} \leq C\delta_T^2.
\end{aligned}
$$

Inequality (I) follows from the definition of the best linear predictor $\boldsymbol{\beta}^{(0)}$, while (II) follows analogously to the steps taken for $A_{1,3}$. (III) follow again from repeated application of the dual norm inequality; take a $p \times n$ matrix A, a $n \times m$ matrix B and an $m \times 1$ vector c, then $\|ABc\|_1 \leq \|c\|_\infty \sum_{i=1}^p \|a_{i\cdot}B\|_1 \leq \|c\|_\infty \sum_{i=1}^p \|a_{i\cdot}\|_\infty \sum_{j=1}^m \|b_{\cdot j}\|_1$. Steps (IV) and (V) then follow from Assumption 4(f-d) and the results for A.

To prove $B_{1,2} = o_p(1)$ we can use the same empirical process bound in 3.64. Then $B_{1,2}$ follows analogously to $B_{1,1,2}$. □

We have proved that post-selected parts A and B in (3.56) are close with high-probability to the same expressions evaluated on the true, fixed dimensional active set $\boldsymbol{S}_0$. It follows that as $\boldsymbol{X}_{\boldsymbol{S}_0} \equiv \boldsymbol{X}_{-GC}$ then the expression for the deviation of the OLS estimator of $\boldsymbol{\phi}^*$ from the true counterpart in (3.56) now becomes:

$$
\boldsymbol{D}_T(\hat{\boldsymbol{\phi}}^* - \boldsymbol{\phi}^*) = \left(\boldsymbol{D}_T^{-1}(\boldsymbol{W}_d^{*\prime}\mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\boldsymbol{W}_d^*)\boldsymbol{D}_T^{-1}\right)^{-1}\boldsymbol{D}_T^{-1}\left(\boldsymbol{W}_d^{*\prime}\mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\boldsymbol{u}\right). \tag{3.67}
$$

Thus, it remains to show how the lag-augmentation provides asymptotic normality of (3.67). To do so, consider the $N_\varphi(p+d) \times N_\varphi(p+d)$ matrix

$$
\begin{aligned}
\left(\boldsymbol{W}_d^{*\prime}\boldsymbol{W}_d^*\right) &= \left((\boldsymbol{W}_{d,A}, \boldsymbol{W}_{d,B}, \boldsymbol{W}_{d,C})'(\boldsymbol{W}_{d,A}, \boldsymbol{W}_{d,B}, \boldsymbol{W}_{d,C})\right) \\
&= \begin{pmatrix} \boldsymbol{W}_{d,A}'\boldsymbol{W}_{d,A} & \boldsymbol{W}_{d,A}'\boldsymbol{W}_{d,B} & \boldsymbol{W}_{d,A}'\boldsymbol{W}_{d,C} \\ \boldsymbol{W}_{d,B}'\boldsymbol{W}_{d,A} & \boldsymbol{W}_{d,B}'\boldsymbol{W}_{d,B} & \boldsymbol{W}_{d,B}'\boldsymbol{W}_{d,C} \\ \boldsymbol{W}_{d,C}'\boldsymbol{W}_{d,A} & \boldsymbol{W}_{d,C}'\boldsymbol{W}_{d,B} & \boldsymbol{W}_{d,C}'\boldsymbol{W}_{d,C} \end{pmatrix},
\end{aligned}
$$

where $\boldsymbol{W}_{d,A}$ is $T \times N_\varphi p$ and $\boldsymbol{W}_{d,B\cup C}$ is $T \times N_\varphi d$. In fact, $\boldsymbol{W}_d^*$ contains the variables interested in the hypothesis at test i.e., the Granger causing and Granger caused transformed in their d-differences plus the augmented lags which either get $d-1$ one or no differences (see Section 3.2). Recall $\boldsymbol{D}_T$ is a $N_\varphi(p+d)\times N_\varphi(p+d)$ diagonal matrix and similarly let the $\hat{s} \times \hat{s}$ matrix

$$\tilde{\boldsymbol{D}}_T := \begin{pmatrix} \sqrt{T} I_A & 0 & 0 \\ 0 & T I_B & 0 \\ 0 & 0 & T^2 I_C \end{pmatrix},$$

where $A + B + C = \hat{s}$ i.e., the scaling depends on the amount of $I(0)$ variables (A), $I(1)$ variables (B) and $I(2)$ variables (C) in the true support $\boldsymbol{S}_0$. Then, the following results applies.

$$\begin{aligned} &\boldsymbol{D}_T^{-1}\left(\boldsymbol{W}_d^{*\prime}\boldsymbol{W}_d\right)\boldsymbol{D}_T^{-1} \xrightarrow{d} \\ &\xrightarrow{d} \begin{pmatrix} \boldsymbol{\Sigma}_A^{\boldsymbol{ww}} & 0 & 0 \\ 0 & \int_0^1 \boldsymbol{B}_B^{\boldsymbol{ww}}(s)'\boldsymbol{B}_B^{\boldsymbol{ww}}(s)ds & \int_0^1 \bar{\boldsymbol{B}}_C^{\boldsymbol{ww}}(s)'\boldsymbol{B}_B^{\boldsymbol{ww}}(s)ds \\ 0 & \int_0^1 \bar{\boldsymbol{B}}_C^{\boldsymbol{ww}}(s)'\boldsymbol{B}_B^{\boldsymbol{ww}}(s)ds & \int_0^1 \bar{\boldsymbol{B}}_C^{\boldsymbol{ww}}(s)'\bar{\boldsymbol{B}}_C^{\boldsymbol{ww}}(s)ds \end{pmatrix} =: \boldsymbol{Q}, \end{aligned} \tag{3.68}$$

by Lemma 3.1 and (g), (m), (f), (i), (l) of Lemma 3.2 and where the notation uses the same superscripts of $\boldsymbol{\Sigma}^j$ also for the vector Brownian motions $\boldsymbol{B}$. To see (3.68) is sufficient to observe that

$$\begin{aligned} &\boldsymbol{D}_T^{-1}\left(\boldsymbol{W}_d^{*\prime}\boldsymbol{W}_d^*\right)\boldsymbol{D}_T^{-1} = \\ &= \begin{pmatrix} T^{-1}\boldsymbol{W}_{d,A}^{'}\boldsymbol{W}_{d,A} & T^{-3/2}\boldsymbol{W}_{d,A}^{'}\boldsymbol{W}_{d,B} & T^{-5/2}\boldsymbol{W}_{d,A}^{'}\boldsymbol{W}_{d,C} \\ T^{-3/2}\boldsymbol{W}_{d,B}^{'}\boldsymbol{W}_{d,A} & T^{-2}\boldsymbol{W}_{d,B}^{'}\boldsymbol{W}_{d,B} & T^{-3}\boldsymbol{W}_{d,B}^{'}\boldsymbol{W}_{d,C} \\ T^{-5/2}\boldsymbol{W}_{d,C}^{'}\boldsymbol{W}_{d,A} & T^{-3}\boldsymbol{W}_{d,C}^{'}\boldsymbol{W}_{d,B} & T^{-4}\boldsymbol{W}_{d,C}^{'}\boldsymbol{W}_{d,C} \end{pmatrix}. \end{aligned}$$

The diagonal elements converge respectively to $\boldsymbol{\Sigma}_A^{\boldsymbol{ww}}$ by Lemma 3.1, $\int_0^1 \boldsymbol{B}_B^{\boldsymbol{ww}}(s)'\boldsymbol{B}_B^{\boldsymbol{ww}}(s)ds$ by result (g) of Lemma 3.2 and $\int_0^1 \bar{\boldsymbol{B}}_C^{\boldsymbol{ww}}(s)'d\boldsymbol{B}_\xi^{\boldsymbol{ww}}(s)$ by result (m) of Lemma 3.2.

The elements (2,3) and (3,2) follow from result (l) of Lemma 3.2. All the remaining outer-diagonal elements refers to result (f),(i),(l) of Lemma 3.2 but since their rate exceeds those of Lemma 3.2, they all converge to 0.

$$\left(\boldsymbol{u}'\boldsymbol{W}_d^*\right)\boldsymbol{D}_T^{-1} \xrightarrow{d} \left(\boldsymbol{\zeta}_1^{\boldsymbol{w}} \quad \int_0^1 \boldsymbol{B}_B^{\boldsymbol{w}}(s)' d\boldsymbol{B}_u^{\boldsymbol{w}}(s) \quad \int_0^1 \bar{\boldsymbol{B}}_C^{\boldsymbol{w}}(s)' d\boldsymbol{B}_u^{\boldsymbol{w}}(s)\right) =: \boldsymbol{H}. \tag{3.69}$$

To see (3.69), similarly to (3.68) we get

$$\left(\boldsymbol{u}'\boldsymbol{W}_d^*\right)\boldsymbol{D}_T^{-1} = \left(T^{-1/2}\boldsymbol{u}'\boldsymbol{W}_{d,A}^* \quad T^{-1}\boldsymbol{u}'\boldsymbol{W}_{d,B}^* \quad T^{-2}\boldsymbol{u}'\boldsymbol{W}_{d,C}^*\right).$$

Elements converge respectively to $\boldsymbol{\zeta}_1^{\boldsymbol{w}}$ for the first column, $\int_0^1 \boldsymbol{B}_B^{\boldsymbol{w}}(s)' d\boldsymbol{B}_\xi^{\boldsymbol{w}}(s)$ for the second and finally to $\int_0^1 \bar{\boldsymbol{B}}_C^{\boldsymbol{w}}(s)' d\boldsymbol{B}_\xi^{\boldsymbol{w}}(s)$ for the third column, respectively by Lemma 3.1 and results (e), (h) of Lemma 3.2.

For the cross-products between $\boldsymbol{X}_{\boldsymbol{S}_0}$ and $\boldsymbol{W}_d^*$ we get

$$\tilde{\boldsymbol{D}}_T^{-1}\left(\boldsymbol{X}_{\boldsymbol{S}_0}'\boldsymbol{W}_d^*\right)\boldsymbol{D}_T^{-1} \xrightarrow{d}$$
$$\xrightarrow{d} \begin{pmatrix} \boldsymbol{\Sigma}_A^{\boldsymbol{wx}} & 0 & 0 \\ 0 & \int_0^1 \boldsymbol{B}_B^{\boldsymbol{wx}}(s)'\boldsymbol{B}_B(s)ds & \int_0^1 \bar{\boldsymbol{B}}_C^{\boldsymbol{wx}}(s)'\boldsymbol{B}_B^{\boldsymbol{wx}}(s)ds \\ 0 & \int_0^1 \bar{\boldsymbol{B}}_C^{\boldsymbol{wx}}(s)'\boldsymbol{B}_B^{\boldsymbol{wx}}(s)ds & \int_0^1 \bar{\boldsymbol{B}}_C^{\boldsymbol{wx}}(s)'\bar{\boldsymbol{B}}_C^{\boldsymbol{wx}}(s)ds \end{pmatrix} =: \boldsymbol{K}. \tag{3.70}$$

(3.70) follows in the exact same way of (3.68) i.e., by result (l), (m) of Lemma 3.2.

$$\tilde{\boldsymbol{D}}_T^{-1}\left(\boldsymbol{X}_{\boldsymbol{S}_0}'\boldsymbol{X}_{\boldsymbol{S}_0}\right)\tilde{\boldsymbol{D}}_T^{-1} \xrightarrow{d}$$
$$\xrightarrow{d} \begin{pmatrix} \boldsymbol{\Sigma}_A^{\boldsymbol{xx}} & 0 & 0 \\ 0 & \int_0^1 \boldsymbol{B}_B^{\boldsymbol{xx}}(s)'\boldsymbol{B}_B^{\boldsymbol{xx}}(s)ds & \int_0^1 \bar{\boldsymbol{B}}_C^{\boldsymbol{xx}}(s)'\boldsymbol{B}_B^{\boldsymbol{xx}}(s)ds \\ 0 & \int_0^1 \bar{\boldsymbol{B}}_C^{\boldsymbol{xx}}(s)'\boldsymbol{B}_B^{\boldsymbol{xx}}(s)ds & \int_0^1 \bar{\boldsymbol{B}}_C^{\boldsymbol{xx}}(s)'\bar{\boldsymbol{B}}_C^{\boldsymbol{xx}}(s)ds \end{pmatrix} =: \boldsymbol{J}, \tag{3.71}$$

and

$$\left(\boldsymbol{u}'\boldsymbol{X}_{\boldsymbol{S}_0}\right)\tilde{\boldsymbol{D}}_T^{-1} \xrightarrow{d} \left(\boldsymbol{\zeta}_1^{\boldsymbol{x}} \quad \int_0^1 \boldsymbol{B}_B^{\boldsymbol{x}}(s)'d\boldsymbol{B}_u^{\boldsymbol{x}}(s) \quad \int_0^1 \bar{\boldsymbol{B}}_C^{\boldsymbol{x}}(s)'d\boldsymbol{B}_u^{\boldsymbol{x}}(s)\right) =: \boldsymbol{R}, \tag{3.72}$$

follow from (m), (h) of Lemma 3.2.

It follows from these results that

$$\begin{aligned}&\boldsymbol{D}_T^{-1}\left(\boldsymbol{W}_d^{*\prime}\mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\boldsymbol{W}_d^*\right)\boldsymbol{D}_T^{-1} = \\ &= \underbrace{\boldsymbol{D}_T^{-1}\left(\boldsymbol{W}_d^{*\prime}\boldsymbol{I}\boldsymbol{W}_d^*\right)\boldsymbol{D}_T^{-1}}_{(i)} - \underbrace{\boldsymbol{D}_T^{-1}\left[\boldsymbol{W}_d^{*\prime}\left(\boldsymbol{X}_{\boldsymbol{S}_0}\left(\boldsymbol{X}_{\boldsymbol{S}_0}'\boldsymbol{X}_{\boldsymbol{S}_0}\right)^{-1}\boldsymbol{X}_{\boldsymbol{S}_0}'\right)\boldsymbol{W}_d^*\right]\boldsymbol{D}_T^{-1}}_{(ii)},\end{aligned}$$

where: $(i) \xrightarrow{d} \boldsymbol{Q}$, $(ii) \xrightarrow{d} \boldsymbol{K}'\boldsymbol{J}^{-1}\boldsymbol{K}$.
Also:

$$\begin{aligned}&\boldsymbol{D}_T^{-1}\left(\boldsymbol{W}_d^{*\prime}\mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\boldsymbol{u}\right) = \\ &= \underbrace{\boldsymbol{D}_T^{-1}\boldsymbol{W}_d^{*\prime}\boldsymbol{I}\boldsymbol{u}}_{(iii)} - \underbrace{\boldsymbol{D}_T^{-1}\left[\boldsymbol{W}_d^{*\prime}\left(\boldsymbol{X}_{\boldsymbol{S}_0}\left(\boldsymbol{X}_{\boldsymbol{S}_0}'\boldsymbol{X}_{\boldsymbol{S}_0}\right)^{-1}\boldsymbol{X}_{\boldsymbol{S}_0}'\right)\boldsymbol{u}\right]}_{(iv)},\end{aligned}$$

where: $(iii) \xrightarrow{d} \boldsymbol{H}'$, $(iv) \xrightarrow{p} \boldsymbol{K}'\boldsymbol{J}^{-1}\boldsymbol{R}'$.

Therefore, we can conclude

$$\begin{aligned}\boldsymbol{D}_T\left(\hat{\boldsymbol{\phi}}^* - \boldsymbol{\phi}^*\right) &= \left(\boldsymbol{D}_T^{-1}\left(\boldsymbol{W}_d^{*\prime}\mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\boldsymbol{W}_d^*\right)\boldsymbol{D}_T^{-1}\right)^{-1}\boldsymbol{D}_T^{-1}\left(\boldsymbol{W}_d^{*\prime}\mathcal{M}(\boldsymbol{S}_{\boldsymbol{S}_0})\boldsymbol{u}\right) \\ &\xrightarrow{d} \left(\boldsymbol{Q} - \boldsymbol{K}'\boldsymbol{J}^{-1}\boldsymbol{K}\right)^{-1}\left(\boldsymbol{H}' - \boldsymbol{K}'\boldsymbol{J}^{-1}\boldsymbol{R}'\right) \\ &\equiv \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{u}} \otimes \boldsymbol{\Sigma}_A\boldsymbol{\Sigma}_J),\end{aligned} \tag{3.73}$$

where $\boldsymbol{\Sigma}_J = \boldsymbol{\Sigma}_A^{\boldsymbol{ww}} - \boldsymbol{\Sigma}_A^{\boldsymbol{ww}}\boldsymbol{\Sigma}_A^{\boldsymbol{wx}-1}\boldsymbol{\Sigma}_A^{\boldsymbol{xx}} - \boldsymbol{\Sigma}_A^{\boldsymbol{wx}}\boldsymbol{\Sigma}_A^{\boldsymbol{xx}-1}\boldsymbol{\Sigma}_A^{\boldsymbol{wx}} + \boldsymbol{\Sigma}_A^{\boldsymbol{wx}}$.

To see (3.73) first observe that

$$
\begin{aligned}
&\boldsymbol{D}_T^{-1}\left(\boldsymbol{W}_d^{*\prime}\mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\boldsymbol{W}_d^{*}\right)\boldsymbol{D}_T^{-1} = \\
&\underbrace{\boldsymbol{D}_T^{-1}\left(\boldsymbol{W}_d^{*\prime}\boldsymbol{I}\boldsymbol{W}_d^{*}\right)\boldsymbol{D}_T^{-1}}_{(i)} - \\
&-\underbrace{\boldsymbol{D}_T^{-1}\left[\boldsymbol{W}_d^{*\prime}\left(\boldsymbol{X}_{\boldsymbol{S}_0}\tilde{\boldsymbol{D}}_T^{-1}\tilde{\boldsymbol{D}}_T\left(\boldsymbol{X}_{\boldsymbol{S}_0}^{\prime}\boldsymbol{X}_{\boldsymbol{S}_0}\right)^{-1}\tilde{\boldsymbol{D}}_T\tilde{\boldsymbol{D}}_T^{-1}\boldsymbol{X}_{\boldsymbol{S}_0}^{\prime}\right)\boldsymbol{W}_2^{*}\right]\boldsymbol{D}_T^{-1}}_{(ii)}.
\end{aligned}
$$

Then $(ii) = (ii.A)(ii.B)(ii.C) =$ $\left(\boldsymbol{D}_T^{-1}\boldsymbol{W}_d^{*\prime}\boldsymbol{X}_{\boldsymbol{S}_0}\tilde{\boldsymbol{D}}_T^{-1}\right)\left(\tilde{\boldsymbol{D}}_T(\boldsymbol{X}_{\boldsymbol{S}_0}^{\prime}\boldsymbol{X}_{\boldsymbol{S}_0})^{-1}\tilde{\boldsymbol{D}}_T\right)\left(\tilde{\boldsymbol{D}}_T^{-1}\boldsymbol{X}_{\boldsymbol{S}_0}^{\prime}\boldsymbol{W}_d^{*}\boldsymbol{D}_T^{-1}\right)$ where by results in (3.70), (3.71) we get $(ii) = \boldsymbol{K}^{\prime}\boldsymbol{J}^{-1}\boldsymbol{K}$.

Similarly for

$$
\begin{aligned}
&\boldsymbol{D}_T^{-1}\left(\boldsymbol{W}_d^{*\prime}\mathcal{M}(\boldsymbol{X}_{\boldsymbol{S}_0})\boldsymbol{u}\right) = \\
&= \underbrace{\boldsymbol{D}_T^{-1}\boldsymbol{W}_d^{*\prime}\boldsymbol{I}\boldsymbol{u}}_{(iii)} - \\
&-\underbrace{\boldsymbol{D}_T^{-1}\left[\boldsymbol{W}_d^{*\prime}\left(\boldsymbol{X}_{\boldsymbol{S}_0}\tilde{\boldsymbol{D}}_T^{-1}\tilde{\boldsymbol{D}}_T\left(\boldsymbol{X}_{\boldsymbol{S}_0}^{\prime}\boldsymbol{X}_{\boldsymbol{S}_0}\right)^{-1}\tilde{\boldsymbol{D}}_T\tilde{\boldsymbol{D}}_T^{-1}\boldsymbol{X}_{\boldsymbol{S}_0}\right)\boldsymbol{u}\right]}_{(iv)}.
\end{aligned}
$$

Then $(iv) = (iv.A)(iv.B)(iv.C) =$ $\left(\boldsymbol{D}_T^{-1}\boldsymbol{W}_d^{*\prime}\boldsymbol{X}_{\boldsymbol{S}_0}\tilde{\boldsymbol{D}}_T^{-1}\right)\left(\tilde{\boldsymbol{D}}_T(\boldsymbol{X}_{\boldsymbol{S}_0}^{\prime}\boldsymbol{X}_{\boldsymbol{S}_0})^{-1}\tilde{\boldsymbol{D}}_T\right)\left(\tilde{\boldsymbol{D}}_T^{-1}\boldsymbol{X}_{\boldsymbol{S}_0}^{\prime}\boldsymbol{u}\right)$ where by results in (3.70), (3.71), (3.72) we get $(iv) = \boldsymbol{K}^{\prime}\boldsymbol{J}^{-1}\boldsymbol{R}^{\prime}$.
Now, putting the pieces together, let us rewrite $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{J}$ as 2×2 block matrices like

$$
\begin{pmatrix} \boldsymbol{\Sigma}_A^i & 0 \\ 0 & \boldsymbol{G}^i \end{pmatrix}
$$

where $\boldsymbol{G}^i := \begin{pmatrix} \int_0^1 \boldsymbol{B}_B^i(s)^{\prime}\boldsymbol{B}_B^i(s)ds & \int_0^1 \bar{\boldsymbol{B}}_C^i(s)^{\prime}\boldsymbol{B}_B^i(s)ds \\ \int_0^1 \bar{\boldsymbol{B}}_C^i(s)^{\prime}\boldsymbol{B}_B^i(s)ds & \int_0^1 \bar{\boldsymbol{B}}_C^i(s)^{\prime}\bar{\boldsymbol{B}}_C^i(s)ds \end{pmatrix}$
and $i = (\boldsymbol{ww}, \boldsymbol{wx}, \boldsymbol{xx})$.

Also, consider the 1×2 matrix $\boldsymbol{H}, \boldsymbol{R} = \begin{pmatrix} \zeta_1^k & \boldsymbol{Q}^k \end{pmatrix}$ where $\boldsymbol{Q}^k := \begin{pmatrix} \int_0^1 \boldsymbol{B}_B^k(s)' d\boldsymbol{B}_u^k(s) & \int_0^1 \bar{\boldsymbol{B}}_C^k(s)' d\boldsymbol{B}_u^k(s) \end{pmatrix}$ for $k = (\boldsymbol{w}, \boldsymbol{x})$. Therefore,

$$\left(\boldsymbol{Q} - \left(\boldsymbol{K}'\boldsymbol{J}^{-1}\boldsymbol{K}\right)\right)^{-1} =$$

$$= \left(\begin{pmatrix} \boldsymbol{\Sigma}_A^{\boldsymbol{ww}} & 0 \\ 0 & \boldsymbol{G}^{\boldsymbol{ww}} \end{pmatrix} - \left(\begin{pmatrix} \boldsymbol{\Sigma}_A^{\boldsymbol{wx}} & 0 \\ 0 & \boldsymbol{G}^{\boldsymbol{wx}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_A^{\boldsymbol{xx}} & 0 \\ 0 & \boldsymbol{G}^{\boldsymbol{xx}} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\Sigma}_A^{\boldsymbol{wx}} & 0 \\ 0 & \boldsymbol{G}^{\boldsymbol{wx}} \end{pmatrix}\right)\right)^{-1}$$

$$= \begin{pmatrix} \boldsymbol{\Sigma}_A^{\boldsymbol{ww}} - \boldsymbol{\Sigma}_A^{\boldsymbol{wx}} \boldsymbol{\Sigma}_A^{\boldsymbol{xx}-1} \boldsymbol{\Sigma}_A^{\boldsymbol{wx}} & 0 \\ 0 & \boldsymbol{G}^{\boldsymbol{ww}} - \boldsymbol{G}^{\boldsymbol{wx}} \boldsymbol{G}^{\boldsymbol{xx}-1} \boldsymbol{G}^{\boldsymbol{wx}} \end{pmatrix}^{-1}.$$

$$\begin{aligned} \left(\boldsymbol{H}' - \boldsymbol{K}'\boldsymbol{J}^{-1}\boldsymbol{R}'\right) &= \begin{pmatrix} \zeta_1^{\boldsymbol{w}} \\ \boldsymbol{Q}^{\boldsymbol{w}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\Sigma}_A^{\boldsymbol{wx}} & 0 \\ 0 & \boldsymbol{G}^{\boldsymbol{wx}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_A^{\boldsymbol{xx}} & 0 \\ 0 & \boldsymbol{G}^{\boldsymbol{xx}} \end{pmatrix}^{-1} \begin{pmatrix} \zeta_1^{\boldsymbol{x}} \\ \boldsymbol{Q}^{\boldsymbol{x}} \end{pmatrix} \\ &= \begin{pmatrix} \zeta_1^{\boldsymbol{w}} - \boldsymbol{\Sigma}_A^{\boldsymbol{wx}} \boldsymbol{\Sigma}_A^{\boldsymbol{xx}-1} \zeta_1^{\boldsymbol{x}} \\ \boldsymbol{Q}^{\boldsymbol{w}} - \boldsymbol{G}^{\boldsymbol{wx}} \boldsymbol{G}^{\boldsymbol{xx}-1} \boldsymbol{Q}^{\boldsymbol{x}} \end{pmatrix}. \end{aligned}$$

Hence,

$$\begin{pmatrix} \boldsymbol{\Sigma}_A^{\boldsymbol{ww}} - \boldsymbol{\Sigma}_A^{\boldsymbol{wx}} \boldsymbol{\Sigma}_A^{\boldsymbol{xx}-1} \boldsymbol{\Sigma}_A^{\boldsymbol{wx}} & 0 \\ 0 & \boldsymbol{G}^{\boldsymbol{ww}} - \boldsymbol{G}^{\boldsymbol{wx}} \boldsymbol{G}^{\boldsymbol{xx}-1} \boldsymbol{G}^{\boldsymbol{wx}} \end{pmatrix}^{-1} \begin{pmatrix} \zeta_1^{\boldsymbol{w}} - \boldsymbol{\Sigma}_A^{\boldsymbol{wx}} \boldsymbol{\Sigma}_A^{\boldsymbol{xx}-1} \zeta_1^{\boldsymbol{x}} \\ \boldsymbol{Q}^{\boldsymbol{w}} - \boldsymbol{G}^{\boldsymbol{wx}} \boldsymbol{G}^{\boldsymbol{xx}-1} \boldsymbol{Q}^{\boldsymbol{x}} \end{pmatrix} =$$

$$\begin{pmatrix} \left(\boldsymbol{\Sigma}_A^{\boldsymbol{ww}} - \boldsymbol{\Sigma}_A^{\boldsymbol{wx}} \boldsymbol{\Sigma}_A^{\boldsymbol{xx}-1} \boldsymbol{\Sigma}_A^{\boldsymbol{wx}}\right)^{-1} \left(\zeta_1^{\boldsymbol{w}} - \boldsymbol{\Sigma}_A^{\boldsymbol{wx}} \boldsymbol{\Sigma}_A^{\boldsymbol{xx}-1} \zeta_1^{\boldsymbol{x}}\right) \\ (\boldsymbol{G}^{\boldsymbol{ww}} - \boldsymbol{G}^{\boldsymbol{wx}} \boldsymbol{G}^{\boldsymbol{xx}-1} \boldsymbol{G}^{\boldsymbol{wx}})^{-1} (\boldsymbol{Q}^{\boldsymbol{w}} - \boldsymbol{G}^{\boldsymbol{wx}} \boldsymbol{G}^{\boldsymbol{xx}-1} \boldsymbol{Q}^{\boldsymbol{x}}) \end{pmatrix}.$$

Since the null hypothesis (3.11) tests only the first $N_\varphi p$ terms, only the first row of the matrices involved here are considered, namely

$$\begin{aligned} &\boldsymbol{\Sigma}_A^{\boldsymbol{ww}-1} \boldsymbol{\zeta}_{\mathbf{1}}^{\boldsymbol{w}} - \boldsymbol{\Sigma}_A^{\boldsymbol{ww}-1} \boldsymbol{\Sigma}_A^{\boldsymbol{wx}} \boldsymbol{\Sigma}_A^{\boldsymbol{xx}-1} \boldsymbol{\zeta}_{\mathbf{1}}^{\boldsymbol{x}} - \boldsymbol{\Sigma}_A^{\boldsymbol{wx}-1} \boldsymbol{\Sigma}_A^{\boldsymbol{xx}} \boldsymbol{\Sigma}_A^{\boldsymbol{wx}-1} \boldsymbol{\zeta}_{\mathbf{1}}^{\boldsymbol{w}} + \\ &+ \boldsymbol{\Sigma}_A^{\boldsymbol{wx}-1} \boldsymbol{\Sigma}_A^{\boldsymbol{xx}} \boldsymbol{\Sigma}_A^{\boldsymbol{wx}-1} \boldsymbol{\Sigma}_A^{\boldsymbol{wx}} \boldsymbol{\Sigma}_A^{\boldsymbol{xx}-1} \boldsymbol{\zeta}_{\mathbf{1}}^{\boldsymbol{x}} = \\ &= \boldsymbol{\Sigma}_A^{\boldsymbol{ww}-1} \boldsymbol{\zeta}_{\mathbf{1}}^{\boldsymbol{w}} - \boldsymbol{\Sigma}_A^{\boldsymbol{ww}-1} \boldsymbol{\Sigma}_A^{\boldsymbol{wx}} \boldsymbol{\Sigma}_A^{\boldsymbol{xx}-1} \boldsymbol{\zeta}_{\mathbf{1}}^{\boldsymbol{x}} - \boldsymbol{\Sigma}_A^{\boldsymbol{wx}-1} \boldsymbol{\Sigma}_A^{\boldsymbol{xx}} \boldsymbol{\Sigma}_A^{\boldsymbol{wx}-1} \boldsymbol{\zeta}_{\mathbf{1}}^{\boldsymbol{w}} + \\ &+ \boldsymbol{\Sigma}_A^{\boldsymbol{wx}-1} \boldsymbol{\zeta}_{\mathbf{1}}^{\boldsymbol{x}}. \end{aligned}$$

Now, since $\boldsymbol{\zeta}_{\mathbf{1}}^k \sim \mathcal{N}(0, \boldsymbol{\Sigma}_A \otimes \boldsymbol{\Sigma}_u)$ then,

$$\left(\boldsymbol{Q}-\boldsymbol{K}'\boldsymbol{J}^{-1}\boldsymbol{K}\right)^{-1}\left(\boldsymbol{H}-\boldsymbol{K}'\boldsymbol{J}^{-1}\boldsymbol{R}'\right)\xrightarrow{d}\mathcal{N}(0,\boldsymbol{\Sigma}_u\otimes\boldsymbol{\Sigma}_A\boldsymbol{\Sigma}_J),$$

where $\boldsymbol{\Sigma}_J=\boldsymbol{\Sigma}_A^{\boldsymbol{ww}}-\boldsymbol{\Sigma}_A^{\boldsymbol{ww}}\boldsymbol{\Sigma}_A^{\boldsymbol{wx}-1}\boldsymbol{\Sigma}_A^{\boldsymbol{xx}}-\boldsymbol{\Sigma}_A^{\boldsymbol{wx}}\boldsymbol{\Sigma}_A^{\boldsymbol{xx}-1}\boldsymbol{\Sigma}_A^{\boldsymbol{wx}}+\boldsymbol{\Sigma}_A^{\boldsymbol{wx}}$ and this concludes the proof.

The results stated in (3.73) guarantees that our null hypothesis in (3.11), which involves only differentiated variables, can be tested in the usual way using χ^2-tests as Wald or Lagrange Multipliers. For the purpose of this chapter we focus on the LM test, note however that the two can be written in the same form with the only difference being that the variance-covariance matrix of the error term for the LM test refers to the restricted least squares instead of the unrestricted ones (see Appendix A). □

Proof of Theorem 3.2. Using the same definitions as in Theorem 3.1,

$$\begin{aligned}
&R^2\equiv\hat{\boldsymbol{\xi}}'\hat{\boldsymbol{\xi}}-\hat{\boldsymbol{\nu}}'\hat{\boldsymbol{\nu}}=\\
&=\boldsymbol{y}'\mathcal{M}(\boldsymbol{X}_{\hat{S}}\cup\boldsymbol{X}_{GC}^{+})\boldsymbol{X}_{GC}\left[\boldsymbol{X}_{GC}'\mathcal{M}(\boldsymbol{X}_{\hat{S}}\cup\boldsymbol{X}_{GC}^{+})\boldsymbol{X}_{GC}\right]^{-1}\\
&\boldsymbol{X}_{GC}'\mathcal{M}(\boldsymbol{X}_{\hat{S}}\cup\boldsymbol{X}_{GC}^{+})\boldsymbol{y}\\
&=A_T'B_T^{-1}A_T=(\boldsymbol{D}_T^{-1}\boldsymbol{e}'\boldsymbol{u})'(\boldsymbol{D}_T^{-1}\boldsymbol{e}'\boldsymbol{e}\boldsymbol{D}_T^{-1})^{-1}(\boldsymbol{D}_T^{-1}\boldsymbol{e}'\boldsymbol{u}),
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\boldsymbol{D}_T^{-1}\hat{\boldsymbol{\xi}}'\hat{\boldsymbol{\xi}}\boldsymbol{D}_T^{-1}=\\
&=\boldsymbol{D}_T^{-1}\boldsymbol{y}'\mathcal{M}(\boldsymbol{X}_{\hat{S}}\cup\boldsymbol{X}_{GC}^{+})\boldsymbol{y}\boldsymbol{D}_T^{-1}=\\
&=\boldsymbol{D}_T^{-1}\boldsymbol{u}'\boldsymbol{u}\boldsymbol{D}_T^{-1}+\underbrace{\boldsymbol{D}_T^{-1}\boldsymbol{\beta}_{GC}'\boldsymbol{X}_{GC}'\mathcal{M}(\boldsymbol{X}_{\hat{S}}\cup\boldsymbol{X}_{GC}^{+})\boldsymbol{X}_{GC}\boldsymbol{\beta}_{GC}\boldsymbol{D}_T^{-1}}_{D_{T,1}}\\
&+\underbrace{\boldsymbol{D}_T^{-1}\boldsymbol{\beta}_{-GC}'\boldsymbol{X}_{-GC}'\mathcal{M}(\boldsymbol{X}_{\hat{S}}\cup\boldsymbol{X}_{GC}^{+})\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC}\boldsymbol{D}_T^{-1}}_{D_{T,2}}+\\
&+\underbrace{2\boldsymbol{D}_T^{-1}\boldsymbol{\beta}_{GC}'\boldsymbol{X}_{GC}'\mathcal{M}(\boldsymbol{X}_{\hat{S}}\cup\boldsymbol{X}_{GC}^{+})\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC}\boldsymbol{D}_T^{-1}}_{D_{T,3}}\\
&+2\underbrace{\boldsymbol{D}_T^{-1}\boldsymbol{u}'\mathcal{M}(\boldsymbol{X}_{\hat{S}}\cup\boldsymbol{X}_{GC}^{+})\boldsymbol{X}_{GC}\boldsymbol{\beta}_{GC}\boldsymbol{D}_T^{-1}}_{D_{T,4}}+
\end{aligned}$$

$$+ 2\underbrace{\boldsymbol{D}_T^{-1}\boldsymbol{u}'\mathcal{M}(\boldsymbol{X}_{\hat{S}} \cup \boldsymbol{X}_{GC}^{+})\boldsymbol{X}_{-GC}\boldsymbol{\beta}_{-GC}\boldsymbol{D}_T^{-1}}_{D_{T,5}}$$
$$- \underbrace{\boldsymbol{D}_T^{-1}\boldsymbol{u}'\mathcal{P}(\boldsymbol{X}_{\hat{S}} \cup \boldsymbol{X}_{GC}^{+})\boldsymbol{u}\boldsymbol{D}_T^{-1}}_{D_{T,6}}.$$

Each term $D_{T,1} - D_{T,6}$ can be proved with a similar strategy to the proof of Theorem 3.1 to be close with high-probability to the true $|\boldsymbol{S}_0|$ sparse model. Then, letting $\boldsymbol{Z}_p \sim N(0, \boldsymbol{I}_p)$, it follows from Assumption 4(b) and the fact that $\boldsymbol{\Omega} = \sigma^2\boldsymbol{\Sigma}_{GC|-GC}$ for $\sigma^2 = \lim_{T\to\infty} \boldsymbol{D}_T^{-1}\mathbb{E}(\boldsymbol{u}'\boldsymbol{u})\boldsymbol{D}_T^{-1}$ that

$$TR^2 \xrightarrow{d} \frac{\boldsymbol{Z}_p'\boldsymbol{\Omega}^{1/2\prime}\boldsymbol{\Sigma}_{GC|-GC}^{-1}\boldsymbol{\Omega}^{1/2}\boldsymbol{Z}_p}{\sigma^2} = \boldsymbol{Z}_p'\boldsymbol{Z}_p = \chi_p^2, \qquad \text{as } T \to \infty. \quad \square$$

Appendix D Additional material

Table 3.9: Simulation results for the PDS-LA-LM Granger causality test using χ^2 distribution

DGP	Size/Power	ρ	K\T	50	100	200	500	1000
1	Size	0	10	25.9	13.8	7.0	4.9	4.2
			20	52.5	16.6	8.4	4.9	5.7
			50	39.5	14.5	12.0	6.0	5.0
			100	41.6	18.8	8.6	6.5	5.1
1	Power	0	10	41.6	52.1	81.0	99.8	100
			20	58.2	49.4	75.1	99.6	100
			50	45.7	46.0	70.4	98.9	100
			100	49.3	45.5	71.2	99.3	100
2	Size	0	10	23.8	12.3	6.8	6.4	5.3
			20	46.6	15.7	10.3	6.0	7.6
			50	34.2	15.6	8.9	7.5	6.7
			100	36.9	17.7	9.3	7.2	6.1
2	Power	0	10	33.2	37.0	56.4	95.5	99.9
			20	48.1	36.4	56.0	93.2	99.9
			50	38.2	34.6	52.4	91.2	99.9
			100	40.9	32.5	54.2	91.9	99.8
1	Size	0.7	10	31.4	14.1	7.3	5.9	5.1
			20	52.6	20.5	10.3	6.3	5.1
			50	37.6	14.9	18.0	8.8	6.2
			100	41.0	20.6	9.3	10.0	5.8
1	Power	0.7	10	36.4	33.9	46.0	86.3	99.4
			20	52.1	36.4	46.0	85.0	99.7
			50	41.7	29.8	47.6	83.0	98.8
			100	41.0	31.1	40.4	78.9	98.4
2	Size	0.7	10	28.9	14.8	7.9	7.1	8.2
			20	51.9	21.0	13.1	7.5	8.1
			50	32.8	17.3	15.8	10.2	8.1
			100	39.5	19.4	12.1	12.3	9.9
2	Power	0.7	10	33.5	28.2	40.7	79.8	98.3
			20	56.0	33.9	41.1	76.6	98.6
			50	42.4	33.4	44.7	75.0	98.4
			100	47.1	36.2	46.3	78.5	98.0

Notes: Size and Power for the different DGPs are reported for 1000 replications. $T = (50, 100, 200, 500)$ is the time series length, $K = (10, 20, 50, 100)$ the number of variables in the system, the lag-length is fixed to $p = 2$ and BIC is used to select the tuning parameter for the lasso. ρ indicates the correlation employed to simulate the time series with the Toeplitz covariance matrix.

Table 3.10: Selection of p, other frequencies, DGP1, $\rho = 0$

	$T = 50$						$T = 100$						$T = 200$						$T = 500$						$T = 1000$					
K	AIC			BIC			AIC			BIC			AIC			BIC			AIC			BIC			AIC			BIC		
	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3
10	0	4	25	0	0	0	0	6	2	0	0	0	0	4	0	0	0	0	0	2	1	0	0	0	0	4	0	0	0	0
20	0	11	6	1	0	0	0	3	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
50	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Notes: the values reported are percentage of incorrectly finding the lag-length p being $= 1, 3, > 3$ out of 100 replications.

Table 3.11: Selection of p, other frequencies, DGP2, $\rho = 0$

	$T = 50$						$T = 100$						$T = 200$						$T = 500$						$T = 1000$					
K	AIC			BIC			AIC			BIC			AIC			BIC			AIC			BIC			AIC			BIC		
	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3
10	0	14	28	4	0	0	0	22	13	0	0	0	0	34	32	0	2	0	0	31	69	0	18	0	0	0	100	0	77	1
20	4	15	13	27	0	0	0	17	9	0	0	0	0	57	31	0	0	0	0	1	99	0	43	0	0	0	100	0	89	11
50	0	11	4	0	0	0	0	16	4	0	0	0	0	52	35	0	0	0	0	0	100	0	55	0	0	0	100	0	87	13
100	0	7	0	0	0	0	0	58	6	0	0	0	0	57	7	0	0	0	0	0	100	0	37	0	0	0	100	0	88	12

Notes: the values reported are percentage of incorrectly finding the lag-length p being $= 1, 3, > 3$ out of 100 replications.

Table 3.12: Selection of p, other frequencies, DGP2, $\rho = 0.7$

	$T = 50$						$T = 100$						$T = 200$						$T = 500$						$T = 1000$					
K	AIC			BIC			AIC			BIC			AIC			BIC			AIC			BIC			AIC			BIC		
	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3
10	4	13	2	33	0	0	0	14	2	0	0	0	0	28	7	0	1	0	0	49	40	0	11	0	0	10	90	0	51	0
20	18	5	1	69	0	0	0	5	0	0	0	0	0	33	4	0	0	0	0	19	80	0	8	0	0	1	99	0	74	2
50	0	7	2	30	0	0	0	1	0	10	0	0	0	5	0	0	0	0	0	30	69	0	0	0	0	0	100	0	94	0
100	0	1	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	57	42	0	0	0	0	0	100	0	99	0

Notes: the values reported are percentage of incorrectly finding the lag-length p being $= 1, 3, > 3$ out of 100 replications.

Table 3.13: Selection of p, other frequencies, DGP2, $\rho = 0$, $\log|\mathbf{\Omega}| \approx \mathrm{tr}\left(\log(\hat{\mathbf{\Omega}})\right)$

	$T = 50$						$T = 100$						$T = 200$						$T = 500$						$T = 1000$					
K	AIC			BIC			AIC			BIC			AIC			BIC			AIC			BIC			AIC			BIC		
	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3	p=1	p=3	p¿3
10	0	15	19	1	0	0	0	23	18	0	2	0	0	36	34	0	3	0	0	25	75	0	24	1	0	0	100	0	86	5
20	0	20	10	0	0	0	0	30	20	0	0	0	0	45	50	0	2	0	0	0	100	0	77	1	0	0	100	0	70	30
50	0	11	4	0	0	0	0	49	10	0	0	0	0	36	64	0	0	0	0	0	100	0	97	0	0	0	100	0	34	66
100	0	7	0	0	0	0	0	58	6	0	0	0	0	31	69	0	1	0	0	0	100	0	100	0	0	0	100	0	16	84

Notes: the values reported are percentage of incorrectly finding the lag-length p being $= 1, 3, > 3$ out of 100 replications.

4
High-Dimensional Granger Causality for Climatic Attribution[1]

[1] This chapter is based on a joint work with Marina Friedrich from Vrije Universiteit Amsterdam and Stephan Smeekes from Maastricht University.

Abstract

In this chapter we test for Granger causality in high-dimensional vector autoregressive models (VARs) to disentangle and interpret the complex causal chains linking radiative forcings and global temperatures. By allowing for high dimensionality in the model we can enrich the information set with all relevant natural and anthropogenic forcing variables to obtain reliable causal relations. These variables have mostly been investigated in an aggregated form or in separate models in the previous literature. Additionally, our framework allows to ignore the order of integration of the variables and to directly estimate the VAR in levels, thus avoiding accumulating biases coming from unit-root and cointegration tests. This is of particular appeal for climate time series which are well known to contain stochastic trends as well as yielding long memory. We are thus able to display the causal networks linking radiative forcings to global temperatures but also to causally connect radiative forcings among themselves, therefore allowing for a careful reconstruction of a timeline of causal effects among forcings. The robustness of our proposed procedure makes it an important tool for policy evaluation in tackling global climate change.

4.1 Introduction

Investigating the climate, its evolution and the factors responsible for its change is a complicated but fundamental task. The 2018 Intergovernmental Panel on Climate Change (IPCC) special report[2] estimates human activity has caused an increase in global warming[3] of approximately 1.0°C above pre-industrial levels (1850–1900). The outlook for the next 10 to 20 years is that the said increase will reach 1.5°C if the current growth rate persists. The alteration of the global temperature has profound impact on human and natural systems. It is therefore clear that in the effort of policymakers in tackling climate change, the assessment of the factors most responsible for igniting the upward global temperature trend is of great relevance. The chemistry and physics describing the interactions between the atmosphere, the oceans, the land surface and the biosphere are nowadays well understood thanks to decades of climate science research. However, building climate models, broadly categorized by increasing complexity as *Energy Balance Models* (EBMs), *General Circulation Models* (GCMs) and *Earth System Models* (ESMs), is still a very complex process. It requires several steps: first, identification and quantification of the Earth processes; second, a coherent systemic mathematical formulation comprising sensible initial conditions and data-driven evolution of the climate forcings and third, the computing power to solve them. In broad terms, EBMs describe temperature variation as a functional response to incoming and outgoing radiation, declined in both natural and athropogenic sources. GCMs subdivide the Earth sphere into a three dimensional grid of predefined cell size (s) where model results pertaining each cell are passed to neighboring cells in order to model the exchange of energy among regions of the Earth in time (t). Smaller (s, t) attains a higher degree of accuracy while rendering the system complexity substantially more

[2] "*Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty*"

[3] Henceforth defined as: increase in combined surface air and sea surface temperatures averaged over the globe and over a 30-year period.

challenging. ESMs are extensions of GCMs including interactive representations of many biogeochemical cycles (Carbon, Sulfur, Ozone). Upon validation (cf. hindcasting[4]), EBMs, GCMs and ESMs are then employed in projecting current climate into the future.

With the aim of helping to refine forecast accuracy of state of the art climate models and enhancing the understanding of the climate variables interplay in time, we propose a discrete time causal identification framework which hinges on high-dimensional time series techniques in order to discover causal links among global temperature and a set of climate variables. Through the modeling lens of a vector autoregressive model (VAR), causality in mean in the sense of Granger (1969) is considered, conditional on an information set composed of various climate variables. Allowing for a large conditional set helps to become more robust against spurious results. Furthermore, in the proposed procedure little to no care is needed towards the time series properties of the variables considered. The data can enter the model in levels and testing for unit roots and cointegration within the VAR is generally not necessary. This avoids the extra uncertainty that usually comes with pre-testing. Our methodology for high-dimensional non-stationary Granger causality testing is based on Chapter 3.

Given the pressing nature of the topics at stake, both the climate science and the climate econometrics literature greatly expanded in recent years. In the remainder of this section, we first present an overview of some recent advances in the climate econometric literature. Second, we briefly review the related literature on Granger causality for climatic attribution and third, we put this chapter in the context of these related papers.

Pretis (2020) established an important equivalence among two-component EBMs and cointegrated VARs, thus justifying their estimation in discrete time through VAR techniques. Such link is important within the framework of measuring the economic impact of climate change

[4] GCMs are tested by running the models in reverse time, backward into the past to check their in-sample accurateness.

(and vice versa) taking into account non-stationarities affecting estimation and inference on EBMs as well as uncertainties about values of physical parameters, which are often assumed constant over time and with covariance structures imposed without statistical or physical basis. Bennedsen et al. (2020) proposes a statistical state-space global carbon budget model entailing atmospheric CO_2 concentrations, anthropogenic emissions, CO_2 uptake by the terrestrial biosphere and CO_2 uptake by the ocean and marine biosphere.

Estrada et al. (2021) investigate the attribution of anthropogenic influences on the observed warming in regional annual temperatures by means of co-trending tests between sets of variables which include radiative forcing of well-mixed greenhouse gases, sum of all natural and anthropogenic radiative forcing as well as the regional temperatures of interest. The existence of a common nonlinear trend in observed regional air surface temperatures is established where anthropogenic forcings play a major role. Results corroborate the fact that the mean temperatures experienced nowadays, even at a regional level, are extreme when compared to the mid 20th century. Coulombe and Göbel (2021) generate long-run forecasts of Arctic's sea ice cover by using a Bayesian de-seasonalized VAR in levels that can flexibly consider interactions among many variables. This allows them to disentangle the effects of feedback loops and external forcings. Impulse response functions are used to show how the Arctic responds to exogenous anomalies. Diebold et al. (2020) propose a dynamic factor model for estimating four leading Arctic sea ice extent indicators. The Kalman smoother is employed to combine estimations of the indicators while averaging out their respective errors. They find the Sea Ice Index to be statistically optimal alone while no gain is achieved by combining this with other indicators, thus confirming the validity of the NASA algorithm based on it.

On the causal investigation between temperature and forcings, several contribution have already been proposed in the literature. Kaufmann and Stern (1997) employ a bivariate Granger causality framework between temperature anomalies from the Southern (S) and Northern

Hemisphere (N), between 1865 and 1994. The test is conditioned step by step to exogenous variables to observe if the Granger causality from South to North vanishes in the presence of natural and/or anthropogenic forcings. To account for possible unit roots, simulated critical values from the test applied to the bivariate VAR with imposed unit roots are used. They detect Granger causality from Southern to Northern temperatures which remains when natural forcings are added. The link vanishes with the inclusion of anthropogenic forcings. The authors conclude that the results are in line with the hypothesis that the South-to-North causal order is generated by the spatio-temporal pattern of anthropogenic emissions of trace gases and sulphate aerosols. A series of papers follows which stay within in the Granger causality framework (Triacca, 2001; Triacca, 2005; Attanasio and Triacca, 2011; Attanasio, Pasini, et al., 2012; Pasini et al., 2012; Triacca et al., 2013). The authors investigate different systems consisting of temperature anomalies or global surface temperature as well as a set of forcings from CO_2, CH_4, N_2O concentrations, total solar irradiance, stratospheric aerosols, cosmic ray intensity, southern-Pacific, decadal-Atlantic or multidecadal oscillation indeces.

In a bivariate Granger causality model, no causality is detected from CO_2 radiative forcings to global surface temperature in Triacca (2005). The authors use an augmented VAR as proposed by Toda and Yamamoto (1995) to account for non-stationarities. Similarly, Triacca et al. (2013) run the same test in a trivariate system, additionally containing different oscillation indices. The authors find Granger causality from greenhouse gas forcings to temperature only under certain system specifications. Non-linear Granger causality is tested via multi-layer feed-forward neural networks in Attanasio and Triacca (2011) finding CO_2 to unidirectionally Granger cause global temperature. Granger causality is found from anthropogenic forcings to temperature anomalies in Attanasio, Pasini, et al. (2012) while evidence of causal decoupling between total solar irradiance and global temperature is found in Pasini et al. (2012). In 2014, Stern and Kaufmann (2014) review their previous 1997 work. Granger causality tests using a Toda and

Yamamoto (1995) augmented VAR is considered. Three different models with three levels of aggregation are employed. Model I aggregates all forcings into one variable (total forcings). Model II uses aggregate variables for anthropogenic and natural radiative forcing while Model III is the most disaggregated letting forcings (GHG, sulfate aerosols, black carbon, volcanic aerosols, solar irradiance) enter the model separately. The total effect of anthropogenic and natural forcing is tested by imposing joint restrictions that exclude all of the anthropogenic or all of the natural forcings. Their findings for Model I show radiative forcing causes temperature but not vice versa. In Model II, natural forcings cause temperature in all scenarios, but anthropogenic forcings cause temperature only when the black carbon forcing is assumed to be zero and the sulfur forcing is assumed to be weak, thus highlighting the uncertainty about the strength of forcings. There is therefore little evidence that temperature causes anthropogenic forcings. Model III results show that GHG and anthropogenic sulfate aerosol cause temperature in all but one of the samples. The authors cannot find a causal effect for black carbon. Volcanic aerosols cause temperature, while solar irradiance does not in most samples. Model III also shows that temperature causes greenhouse gases. Results indicate that temperature causes carbon dioxide and methane, but temperature has no causal effect on the other non-temperature sensitive greenhouse gases.

Causal discoveries are a delicate matter. In fact, in many cases the above mentioned literature on causality might simply reflect a predictability exercise which surely is of practical interest but it lacks robustness to spurious discoveries. As the climate is a complex system governed by countless sources, it is likely impossible to condition the causal relation on all existing sources. Also, as the number of variables increases, we face two main challenges. On the one hand, the VAR system becomes hugely parametrized and on the other hand, interpretability becomes increasingly challenging. Our approach cannot solve the issue of conditioning the causal relation on all the variables in the climate system as this would not be feasible in general. However, it manages to get closer to this idealized setting. In fact, the active dimensionality reduction through sparsity inducing techniques in the proposed Granger causality test is able to resolve the parametrization issue of the VAR and hence facilitating interpretation of the results. This allows for considering increasingly larger models, possibly very complex and articulated and still attain interpretable results. Such results are fundamental for *climate change attribution* i.e., attributing the detected significant climate change to specific causes.

The remainder of the chapter is organized as follows. Section 4.2 introduces the topic of climate change, its challenges and poses the problem of its attribution. Section 4.3 discusses the methodology of Granger causality testing in high-dimensional levels VARs. Section 4.4 begins the empirical analysis. This is subdivided in three sub-sections: Section 4.4.1 uses aggregated greenhouse gases, similarly Section 4.4.2 where additional crucial variables are added; Section 4.4.3 disaggregates the greenhouse gas series into its three main gas components. Section 4.5 performs a sensitivity analysis first on unit root testing and further on the lag-length specification. Finally, Section 4.6 concludes.

4.2 Climate Change and its Attribution

Given an area of the world, or the entire world itself, its observed average weather over a long span of time is what we commonly refer to as *climate*. In order to describe it, information on many variables of geophysical kind is needed e.g., its average temperature throughout the seasons, the observed pressure, amount of sunshine, rainfall, winds and extreme events like hurricanes and eruptions and many more. It is indeed a large, complex, time varying system of inter-playing variables. Long term systematic changes in the statistics of such variables is what is referred to as *climate change*. If this change would only be naturally induced i.e., caused by changes in exogenous forcings[5] like

[5] A climate forcing is an imposed change in Earth's energy balance, measured in Wm^{-2}. For example, Earth absorbs about 240 Wm^{-2} of solar energy, so if the Sun's brightness increases 1% it is a forcing of +2.4 Wm^{-2}.

e.g., terrestrial orbit, solar emission, aerosols and many other natural internal processes of the climate system, then its study would just be prerogative of geophysicists and scientists in general, trying to explain and forecast consequences, but would less be a matter of public and political debate. The reality of the matter is that surely the climate variation is partly naturally induced but is the human (*anthropogenic*) "fingerprint" that moves the balance needle. Thanks to climate models simulations, it has been widely assessed the extreme unlikeliness that natural variability and natural forcings of the Earth's climate could produce the unprecedented temperature records as observed since mid nineteenth century (see e.g. the IPCC 2007, Solomon et al., 2007). Instead, the sharp upward temperature change observed in the last few decades has to be explained by the anthropogenic emissions. The "good news" is that anthropogenic emissions are an endogenous element in the system and not something we suffer without any remedy possible: we, as humans, emit these gases as a result of many different industrial and every day life processes. Thus, by knowing the source, we should be more in control of our "climate destiny". On paper is all trivial, but the real, practical challenges are of course multiple. First, identification of the sources: many compose the human climate footprint, some are well known like the burning of fossil fuels which with the Industrial Era became the dominant source of anthropogenic emissions; some others are proportionally less media-covered, like deforestation and other land-use changes. Not only is important to identify which emissions are anthropogenic but also crucially *which ones*, and *to what degree*, they are responsible for sistematically changing the climate. The "*which ones*" question is the *attribution* of climate change and this is what we pursue in this work from a statistical perspective. The "*to what degree*" question requires a proper assessment i.e., a quantification of anthropogenic greenhouse gases emissions in the atmosphere and CO_2 especially: the so-called *global carbon budget* (see Friedlingstein et al., 2020; Bennedsen et al., 2020). Greenhouse gases do not just enter the atmosphere to stay but they gets redistributed among the atmosphere, ocean, and terrestrial biosphere: this is referred to as the *global carbon cycle*. The assessment of ongoing and paleo-temperature change is also an important matter in order to define limits of anthropogenic effects on climate. Especially paleoclimate is useful for characterizing long-term ice sheet and sea level response to temperature change (see Hansen, Sato, Kharecha, et al., 2017). Therefore, climate change attribution coupled with a statistical understanding of the global carbon budget and of current and past temperatures is crucial to better understand the global carbon cycle as well as project future climate change and support the development of climate policies. The latter point is indeed the next challenge: policymakers need to be supported from science and act accordingly. Scientists should seek clarity as any apparent inconsistency gets wildly amplified by politics and media. One issue lies of course in the fact that despite the evidence of the anthropogenic climate impact, lots of uncertainties remain in precisely quantifying its influence. As observed in Schneider and Kuntz-Duriseti (2002), the atmosphere of the Earth is missing a suitable control. In other words, we do not know an "undisturbed Earth" that could provide us a reference to compare the current anthropogenic contribution to climate change. Therefore we need to use and rely on models and their climate simulations to estimate how the climate on Earth might have evolved without the human footprint. Models and simulations carry uncertainties that cannot be avoided. Skepticism is however not justifiable, as long as trust in science is put. The concentration of carbon dioxide (CO2) in the atmosphere at the beginning of the Industrial Era in 1750 was approximately 277 parts per million (ppm) (Joos and Spahni, 2008). This concentration became roughly 409.85 ppm in 2019 (Dlugokencky and Tans, 2018). Hansen, Sato, Kharecha, et al. (2017) showed how global temperature has risen out of the Holocene range leaving Earth these days to be as warm as it was during the prior interglacial period[6]. Earth energy imbalance is evident, thus implying more warming is yet to come. However the growth rate of greenhouse gas climate forcing has not decreased but ac-

[6] Hansen, Sato, Kharecha, et al. (2017) observes how the Holocene, which lasts now for over 11 700 years, had relatively stable climate up until the unprecedented warming in the past half century. The Eemian, which lasted from about 130 000 to 115 000 years ago, was instead slightly warmer than the Holocene and this was sufficient to have sea level rise to 6–9 meters greater than today.

celerated in the past decade. Nowadays the global warming rate, based on a linear fit for 1970–2017 (see Hansen, Sato, Kharecha, et al. (2017) Fig. 2b) is +0.18 degree Celsius per decade. The period starting with 1970 sees the highest growth rate of greenhouse gas climate forcing, which has been maintained ever since at approximately +0.4 Wm^{-2} per decade. Is precisely this rate of added climate forcing that needs to be tackled down by policymakers to avoid the already observed adverse climate impacts such as extreme events.

This chapter contributes to the outlined vast global picture of climate change research focusing on climate change attribution. In the following sections we are going to define the statistical modeling framework proposed here and put it into practice with global annual data.

4.3 Methodology: Granger Causality

Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$ be a K -dimensional multiple time series process, where $\boldsymbol{y}_t = (y_{1,t}, \ldots, y_{K,t})'$ is generated by a VAR(p) process

$$\boldsymbol{y}_t = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t, \quad t = p+1, \ldots, T \tag{4.1}$$

where for notational simplicity we assume the variables have zero mean; if not they can be demeaned prior to the analysis, or equivalently a vector of intercepts is added. $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p$ are $K \times K$ parameter matrices and $\boldsymbol{u}_t$ is a martingale difference sequence of error terms. $\boldsymbol{y}_t$ is a K-dimensional vector containing climate time series of length T. We allow K to be large and potentially larger than T. VARs are especially keen to such high-dimensionality since the parameters to estimate grow quadratically with the number of series included. If on the one hand the model dimensionality needs to be handled in order to be able to estimate the parameters, on the other hand one would usually need to take good care of the properties of the time series entering the model before putting the latter to work. In fact, as observed in Chapter 3, standard asymptotic theory is in general not applicable to hypothesis testing in levels VARs if the variables are integrated $I(d)$ or cointegrated $CI(b, d)$

(Park and Phillips, 1988; Sims et al., 1990) given the non-standard limiting distributions of the estimators. One then usually needs to perform statistical unit root tests to check the order of integration of every time series as well as a cointegration rank test to assess the long run relationships among them. When the dimension of a VAR starts to be substantial, testing for unit root all K variables in $\boldsymbol{y}_t$ might induce biases. Integration and cointegration tests in fact suffer from lack of power and they heaviliy depend on the exact specification of the model. Even when one has tested for unit root all those series and corrected for multiple testing and lack of power, the usual next step is taking the differences of the time series in object, where the difference order corresponds to the tested order of integration d. Although this is the praxis and it is a convenient transformation for asymptotic reasons, often is not an innocuous one. Most climate time series definitely do not appear to be stationary in their original levels but they are characterized by stochastic trends (see e.g. Figure 4.1). Taking the d-th difference of the series allows to stabilise the mean by removing changes in the levels and thereby eliminating (or reducing) trends and/or seasonality. However, this induces a loss of information since the long-term memory of the series gets wiped out by the differentiation. Climate time series, among others, are well known to exhibit long memory and hence avoidance of such transformations can be beneficial for the robustness of the final result (see the sensitivity analysis in Section 4.5).

To avoid pre-test biases from unit root and cointegration we follow the procedure outlined in Chapter 3. There, a lag-length augmentation, restricted to the sole variables of interest for testing Granger causality, is employed in a similar vein to Toda and Yamamoto (1995). The adaptive augmentation, as opposed to the lag-augmentation of the full set of K variables, makes it an important extension for the high-dimensional setting. As discussed in Chapter 2 and 3, refitting the selected model with ordinary least squares and hence perform standard Wald-like inference on the coefficient vector of the Granger causing variables is not optimal for performing inference (see the critique of Leeb and Pötscher, 2005). Post selection estimators are not able to converge uniformly in the parameter space to the normal distribution, but only point-wise. Thus a regularization bias will occur in post-selection estimators that will compromise the coverage of any confidence interval. To overcome such issues we focus on the framework developed in Chapter 2 and 3, namely the post-double-selection (PDS) technique. The double variable selection step of both the outcome and the treatment variable on all the controls in the PDS substantially diminishes the omitted variable bias and ensures the errors of the final model are (close enough to) orthogonal with respect to the treatment. Chapter 2 extended the PDS to dependent processes, specifically stationary time series, using $p + 1$ steps where the response variable becomes the Granger caused, the treatment variable becomes the p lags of the Granger causing and finally the controls are all the other variables in the information set. Successively, Chapter 3 extended the PDS Algorithm in Chapter 2 to unit-root non-stationary VARs. The fundamental extra step is employing a lag-augmentation for the post-selection least-square estimator as described in Chapter 3, thus assuring the Granger causing and Granger caused variables to be stationary. We refer to Chapter 3 for a complete treatment of the post-double selection, lag augmented, Lagrange multiplier test (PDS-LA-LM). We are now going to give a series of remarks on its Algorithm 3. The full algorithm is reported in Chapter 3.

Remark 4.1. Algorithm 3 estimates Granger causality directly in levels, disregarding the integration and cointegration properties of the time series in the model. In the analysis in Section 4.4 we compare the same procedure for testing Granger causality with and without lag augmentation. In the case without the augmentation, we first test for unit roots all the variables in the system and we take the appropriate differences of all the series. We show how this is not an innocuous transformation and the results differ, finding less causal discoveries than in the levels analysis as these have been masked by taking the differences of the series.

Remark 4.2. The causal identification in Algorithm 3 still remains anchored to the information set at hand and, as such, truly causal discoveries need to be argued with care. However, the benefit of the methodology developed in Chapter 3 and outlined here is that it allows for high-dimensional information sets. Conditioning is crucial for causal discoveries, to avoid finding spurious causal connections that only reflect predictability.

Remark 4.3. To obtain a data-driven estimate of the lag-length p, we use the modified BIC criterion developed in Chapter 3. Simulations in Chapter 3 show how this procedure to estimate the lag-length works well in practice: being an upper bound, at most it overestimates the lag-length but it does not under-estimate it. The diagonal AR(p) structure estimation is able to circumvent the dimensionality issue: one would otherwise have to estimate larger systems the larger the lag-length and this would blow up tremendously the dimensions even if one would rely on sparsity inducing techniques as the lasso. In fact, even though the lasso should shrink and eventually set to zero the coefficients of those irrelevant parameters, as selection consistency relies on how one tunes the penalty parameter, embracing the high-dimensionality for the purpose of estimating the lag-length only shifts the model selection problem from a BIC applied over estimated AR(p) residuals as in Chapter 3 to the choice of the tuning parameter in the regularization technique. Therefore, lasso might end up erratically estimating a (much) high(er) lag-length than necessary, thus rendering the whole analysis hugely parametrized. One interesting exception is the hierarchical penalties of Nicholson, Wilms, et al. (2020), these include the notion of lag selection into a convex regularizer and they can be used on a set of values for p, possibly varying the lag-length over different variables. However, we do not report a comparison of their performances in this work.

While the proposed lag-length estimation works well in high-dimensional systems, it cannot take into account the type of variables the practitioner is dealing with. Specifically, the procedure cannot a priori recognize if the variables have truly a slow dynamic response. As this could be the case for climate time series, in Section 4.5 we manually augment the estimated lag-length, thus using the latter as lower bound reference. Especially for the identification of causal relations between temperature and greenhouse gases we find it is beneficial to account for quite a larger lag-length.

Remark 4.4. As mentioned in Section 4.3, for the choice of the tuning parameter we rely on the results of Chapter 2, namely BIC is minimized to select the penalty parameter for all the lasso regressions in Algorithm 3. Minimizing information criteria is fast and delivers robust results in practice as evident from several simulation exercises in the literature. A comparison of other methods to tune the penalty parameter is reported in Chapter 2.

4.4 Analysis

4.4.1 Aggregated Greenhouse Gases Analysis (a)

We make use of annual time series data spanning the period from 1850 to 2019. The variables considered, their measurement unit and the source are reported below:

I. **S**: Solar Activity Fe(W/m^2), (Fe: *effective forcings*). Source: Hansen, Sato, Kharecha, et al. (2017).

II. **V**: Stratospheric Aerosols from Volcanic Activity Fe(W/m^2). Source: Hansen, Sato, Kharecha, et al. (2017).

III. **Y**: GDP (log 2010 US$). Source: Maddison Project Database 2020 (Bolt and Zanden (2013)) for $1850 - 1959$, World Bank data for $1960 - 2019$.

IV. **G**: Greenhouse gas concentration Fe(W/m^2). Source: Hansen, Sato, Kharecha, et al. (2017).

V. **A**: Tropospheric Aerosols and Surface Albedo Fe(W/m^2). Source: Hansen, Sato, Kharecha, et al. (2017).

VI. **T**: Temperature Anomaly (°C). Source: Morice et al. (2020).

Figure 4.1 displays the plots of the time series in object.

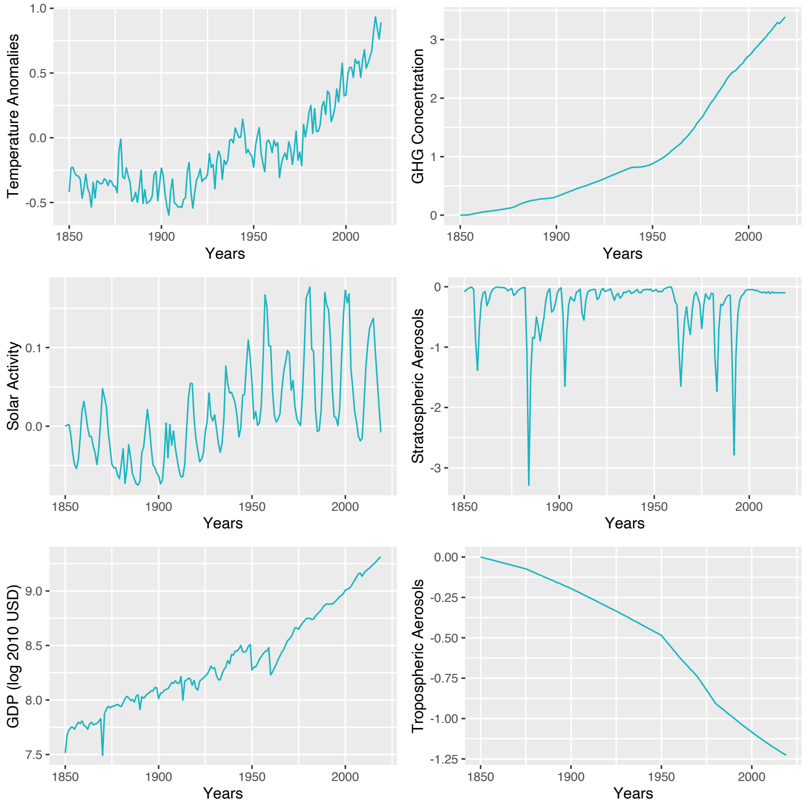


Figure 4.1: Climate Time Series

Remark 4.5. This set of variables considers key factors of the climate evolution. *Effective forcings* of S, V, G, A are considered according to Hansen, Sato, Kharecha, et al. (2017), thus removing the effect of rapid adjustments occurring in the atmosphere which do not relate with longer term surface temperature response. First, we consider temperature anomalies, which is the natural target to be explained in climate change models. From its displayed time series in Figure 4.1 a pronounced increase after 1900 is clearly visible. Second, greenhouse gases. They are partly naturally occurring in the atmosphere and partly are amplified by human activities. As the light energy from the sun penetrates the atmosphere, the Earth absorbs it, thus warming its surface and reemitting some of this energy back as infrared-radiation. The latter is in turn mostly absorbed by the greenhouse gases in the atmosphere which radiate this heat in several directions: partly into the space and partly back to the Earth again, thus contributing in increasing its warming. The *natural greenhouse gas effect* is what allows the Earth's surface temperature to raise up to 33 degree Celsius and hence allows the life on Earth. Anthropogenic emissions due to human activities add to the atmospheric concentrations of the natural greenhouse gases as well as introduce several others that in nature do not occur. This has the effect of increasing the greenhouse gas effect which in turn is responsible of increasing the temperature. Third, aerosols: both in the troposphere and in the stratosphere, they are tiny droplets which have a cooling effect on temperature as they scatter sunlight back to space impeding it to arrive to the Earth surface and warm it. Aerosols can also be distinguished by being of natural and of anthropogenic source. The former comes from e.g., volcanic eruptions, evaporation of seawater, hydrocarbon emissions from forested areas. The latter comes mostly from fuel combustion (diesel and biomass burning produce *black aerosols* absorbing sun's energy) and burning of high-sulfur coal. Fourth, solar activity has an important, purely natural, influence on the climate system. In fact, long term solar variations from both variability in the sun and from the Earth's orbit do affect the climate in thousands of years time. Finally, we also include world GDP. This is of course an objective measure of the amount of value produced and is directly linked to the human activity employed, and hence its emissions.

Remark 4.6. Climate change caused by alterations, or shocks, in forcing can have significant effects on many different processes: atmospheric, geological, biological, oceanographic, chemical, economic. However, these effects can in turn further change the climate, thus creating a *feedback effect*. These can either amplify the initial cause if the direction of the two is concordant (*positive feedback*) or reduce it if their direction is opposite (*negative feedback*). Examples of positive feedbacks are: decrease of the planet's albedo as a consequence of melting of mountain glaciers due to rising temperature. As the melting of ice exposes the less reflective surface land, the albedo i.e., the planet reflectance of solar radiation, will in turn decrease giving way to more solar energy absorption which itself causes additional warming. Similarly, as rising temperature increases evaporation of waters from oceans and lakes, such vapor is itself a greenhouse gas, hence in turn it amplifies warming creating a positive feedback. On the other hand, as the vapor increases its concentration in the atmosphere, this can cause more cloudiness. In turn the clouds raise the Earth's albedo given the increased reflection of the solar radiation. Less reflection implies less energy gets absorbed by the Earth thus decreasing its temperature. These are only few examples of the various feedback effects in climate systems. In Section 4.4.2 and 4.4.3 we highlight these estimated causal feedbacks and cycles among the considered variables.

Let us stack the variables S, V, Y, G, A, T in a VAR model as in (4.1) and we use Algorithm 3 recursively on each of such series in order to obtain the causal network displayed[7] in Figure 4.2.

Figure 4.3 reports a heat-map for the strength of the p-values of the PDS-TY-LM test for the different combinations among the considered series. The darker the color, the smaller the p-value. All white boxes are p-values larger than 15%. Similarly, Figure 4.4 reports a heat-map for the magnitude of the estimated coefficients of the Granger causing variable at the post-selection step, thus accounting for the variables selected via the lasso double-selection. More precisely, we report the sum of the p estimated lag-coefficients of the Granger causing variable where the values are rounded to the fourth decimal and the colours reflect the magnitude in absolute value of such coefficients, where the more red the higher the magnitude.

[7] All the analyses reported in the following sections have been carried out using *R* (R Core Team, 2020). R scripts are available within the package *HDGCvar* which can be downloaded from the Github page of the author: `https://github.com/Marga8/HDGCvar`.

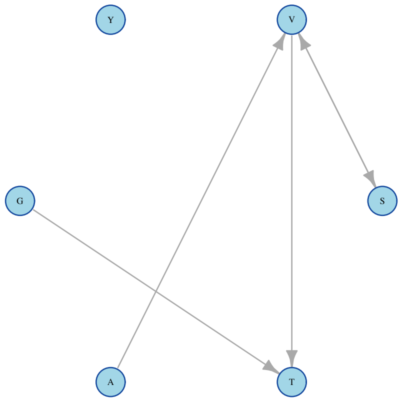


Figure 4.2: Climate Network, $\alpha = 0.1$, $p = 3$

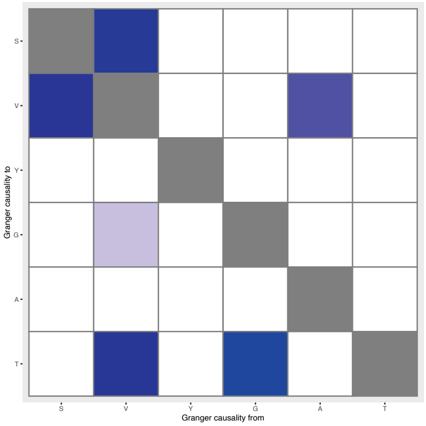


Figure 4.3: P-values heat-map

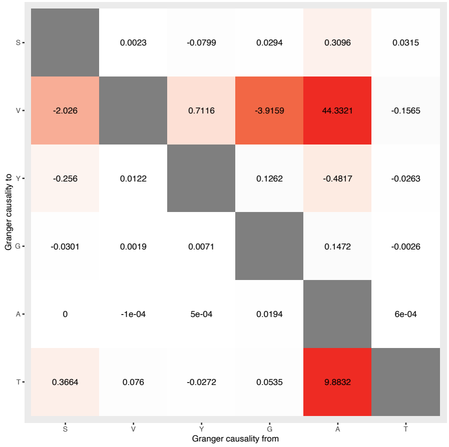


Figure 4.4: Magnitude of Coefficients

We observe a total of 5 direct connections at 10% significance level, namely: greenhouse gas is found to lead temperature (magnitude 0.05) and similarly does stratospheric aerosols (magnitude 0.07). Both connections are in line with the climate literature. As explained in Remark 4.5, among the climate forcings alterations caused by anthropogenic emissions, it is well established how greenhouse gases warm the planet, while at the same time aerosols have a cooling effect (see among others: Mitchell et al., 1995). Tropospheric aerosols & surface albedo (A) are found to lead their Stratospheric counterpart (V), which is sensible given that the Troposphere is the first, lowest layer of the atmosphere ($\approx 0 - 10$ Km), immediately followed by the Stratosphere ($\approx 11 - 50$ Km). Among them an exchange of mass and chemical species is well known to occur (see e.g. Holton et al., 1995). It immediately follows from the said connections that an indirect causal effect occurs between tropospheric aerosols & surface albedo and temperature, passing through stratospheric aerosols. Finally, a feedback causal effect is found between stratospheric aerosols and solar activity (S). This can also be explained: aerosols tend to scatter the sunlight such that some of it gets lost in space without heating the Earth. This process, however, increases the albedo of the planet thus cooling it down. This also explains the indirect connection found among solar activity and temperature (T).

4.4.2 Aggregated Greenhouse Gases Analysis (b)

The dataset I-VI considered in Section 4.4.1 does not consider two important climate variables: the El Niño–Southern Oscillation index (ENSO) and the Ocean heat content (OHC). The former is a standardized index based on the observed sea level pressure differences between Tahiti and Darwin, Australia. It measures the large-scale fluctuations in air pressure occurring between the western and eastern tropical Pacific during El Niño and La Niña episodes.[8] Oscillations of annual CO_2 growth are correlated with global temperature and with the El Niño/La

[8] For the calculation of ENSO we refer to https://www.ncdc.noaa.gov/teleconnections/enso/indicators/soi/

Niña cycle (see Hansen, Sato, Kharecha, et al., 2017). Thompson et al. (2008) shows that El Niño is a natural source of variability responsible for the global warmth of 1939–1945 and strong El Niño events have also occurred in 1997–1998 and 2015–2016 that might have boosted the temperature. On the other hand, OHC also needs to be accounted for in the analysis. In fact, the ocean has lots of thermal inertia and it might take up to centuries before the Earth surface temperature reaches most of its fast-feedback response to a change in climate forcing (Hansen, Russell, et al., 1985). These two series are therefore added to the dataset and the analysis is replicated:

VII. **N**: El Niño–Southern Oscillation index (ENSO) (°C). Source: Climate Research Unit (CRU) at the University of East Anglia[9]

VIII. **O**: Ocean Heat Content (10^{21}J, full depth). Source: Zanna et al. (2019)[10].

As **N** is available from 1866 to 2019 and **O** is available from 1871 to 2018, we trim the full sample from 1871 to 2018, Figure 4.5 displays a plot of the two new variables.

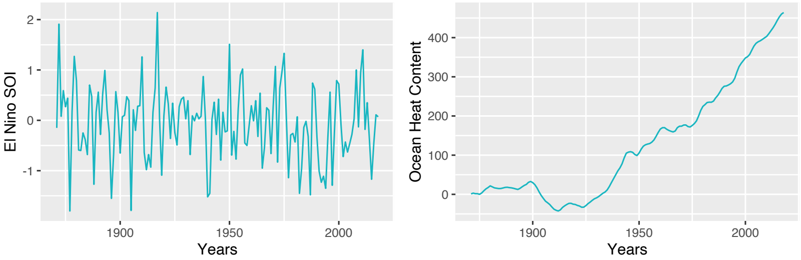


Figure 4.5: Climate Time Series

Figure 4.6 plots the Granger causal network including the variables described above where, as in Section 4.4.1, the arrows represent a (direc-

[9] https://crudata.uea.ac.uk/cru/data/soi/ SOI calculations are based on the method given by Ropelewski and Jones (1987).

[10] https://laurezanna.github.io/post/ohc_pnas_dataset/

tional) causal link that was found at a significance level of 10%. Figure 4.7 presents the corresponding heat map of p-values and Figure 4.8 the heat-map for the coefficient magnitude. The results of this extended causal network give us some key insights. Conditional on all other variables, we find evidence that greenhouse gas forcings (G) Granger causes temperature anomalies (T) (magnitude 1.47). This means that given our included natural forcing as well as production variables, anthropogenic forcings from Greenhouse gases Granger cause temperature. However, temperature does not Granger cause anthropogenic forcings in the same system. This supports the findings of Stern and Kaufmann (2014) who also find a unidirectional link from anthropogenic forcings to temperature in the presence of natural forcings. We additionally find a strong link to temperature from the following variables: GDP (Y), stratospheric aerosols (V), ENSO (N) and ocean heat content (O). The link between stratospheric aerosols and temperature has also been found in Stern and Kaufmann (2014), while ocean heat content is taken as a purely exogenous variable in this chapter. Our results indicate that part of the effect of Greenhouse gas forcings on temperature might run through the ocean which takes up part of the increase in heat. A causal link between ENSO and temperature has not been tested before in the literature. However, Stern and Kaufmann (2014) discuss the option of removing the effect of ENSO on the temperature and modelling the adjusted series. The authors, however, argue that these oscillations are an endogenous part of the climate system and therefore should not be removed. We therefore explicitly consider ENSO in our model and we find a causal link to temperature which runs in both directions. This is not surprising given the previous discussion. In addition, we find that production has a strong effect on Ocean Heat Content and that Solar activity causes Stratospheric aerosols.

From Figure 4.6 we observe two *feedback* relations: stratospheric aerosols with solar activity (also observed in the earlier analysis in Section 4.4.1) and temperature with ENSO. The latter is of particular interest as observed in Houghton et al. (2001). An El Niño event is characterized by positive temperature anomalies in the eastern equatorial Pacific. This reduces the sea surface temperature difference across the tropical Pacific. As a consequence to this, the trade winds from east to west near the equator are weakened and the Southern Oscillation Index becomes anomalously negative, letting sea level to fall in the west and rise in the east by almost 25 cm. At the same time, these weakened winds reduce the rise of cold water in the eastern equatorial Pacific, thereby strengthening the initial positive temperature anomaly. Thus, ENSO influences tropical climate but also possess a global influence: during and following El Niño, the global mean surface temperature increases as the ocean transfers heat to the atmosphere (see Sun and Trenberth, 1998).

From Figure 4.6 we can also identify *cycles*. Greenhouse gases lead OHC which in turn leads stratospheric aerosols which is itself found to feedback to greenhouse gases. There is also an outsource on solar activity which is caused by stratospheric aerosols and it feedbacks to it. The connection between G and O is easily understood: as greenhouse gases act as a blocking layer trapping more energy from the sun, the oceans are absorbing more heat as a consequence and this results in an increase in sea surface temperatures and rise of sea level. The connection between O and V is also justified. Even though we do not find a direct feedback of V on O, there is a cyclic relation among the two outsourced by G. Church et al. (2005) observed how large volcanic eruptions, emitting aerosols in large quantity, result in rapid reductions in ocean heat content and global mean sea level. They bring the example of the eruption of Mount Pinatubo, a stratovolcano in the Philippines, estimating as a consequence of the eruption a reduction in ocean heat content of about $3^{*}10^{22}$J and a global sea-level fall of about 5 mm. Over the three years (coinciding with our lag-length) following such an eruption, the estimated decrease in evaporation is of up to 0.1 mm day^{-1}.

Focusing on the two arguably most interesting connections – Greenhouse gases to temperature and production to temperature – we also plot all possible causal paths for these two relationships in Figures 4.9 and 4.10. These show that there is not only a directional link but the effect can also be indirect by going through (multiple) other variables.

For greenhouse gases we identify four *simple*[11] causal paths: one direct and three indirect, running through the nodes of: ocean heat content, ENSO and stratospheric aerosols. Interestingly, we do not observe a causal path from greenhouse gas passing through GDP. For GDP instead we observe a total of five causal paths among which one direct and four indirect passing through the same nodes as for greenhouse gases but, interestingly, also passing through greenhouse gas itself. The fact that we find a stream of causality from GDP to greenhouse gas but not vice-versa is probably not entirely surprising as production is surely a source of a variety of emissions. More surprisingly, we do not observe an indirect causal stream from greenhouse gases through GDP.

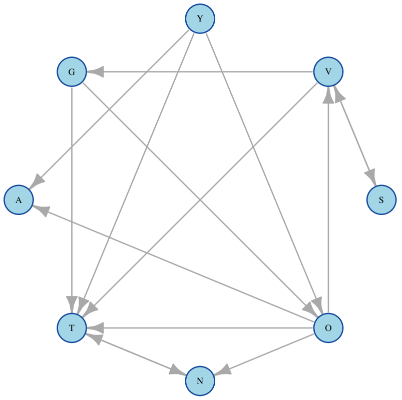


Figure 4.6: Climate Network, $\alpha = 0.1$, $p = 3$

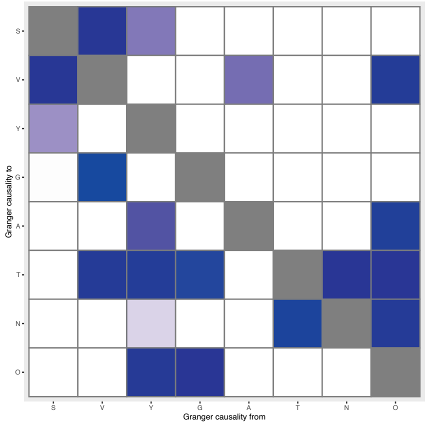
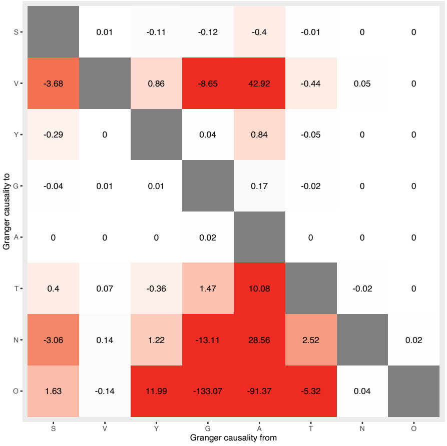


Figure 4.7: P-values heat-map

[11] In graph theory a path is simple if the vertices it visits are not visited more than once.



	S	V	Y	G	A	T	N	O
S		0.01	-0.11	-0.12	-0.4	-0.01	0	0
V	-3.68		0.86	-8.65	42.92	-0.44	0.05	0
Y	-0.29	0		0.04	0.84	-0.05	0	0
G	-0.04	0.01	0.01		0.17	-0.02	0	0
A	0	0	0	0.02		0	0	0
T	0.4	0.07	-0.36	1.47	10.08		-0.02	0
N	-3.06	0.14	1.22	-13.11	28.56	2.52		0.02
O	1.63	-0.14	11.99	-133.07	-91.37	-5.32	0.04	

Figure 4.8: Magnitude of Coefficients

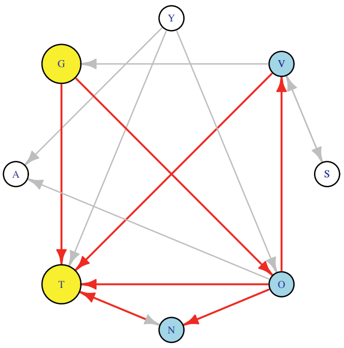


Figure 4.9: G to T paths (4)

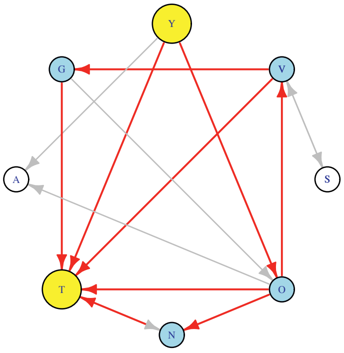


Figure 4.10: Y to T paths (5)

4.4.3 Disaggregated Greenhouse Gases Analysis

In the previous setting we considered an aggregated measure of Greenhouse gas (G) which is meant to represent the combined forcings of all the different anthropogenic emissions. Greenhouse gases (GHGs) increase global Earth warming by absorbing energy and slowing the rate at which the energy escapes to space acting like a barrier-layer around the globe. Therefore, they are divided by their ability to absorb energy (*radiative efficiency*, Wm^{-2} per ppb) and how long they stay in the atmosphere (*lifetime* in years). The three main GHGs for global warming potential are carbon dioxide (CO_2), methane (CH_4), Nitrous Oxide (N_2O). Many others are categorized into: Hydrofluorocarbons, Chlorofluorocarbons, Bromocarbons, Hydrobromocarbons and Halons, Hydrochlorofluorocarbons. We use the historical reconstruction of CO_2, CH_4 and N_2O computed in Meinshausen et al. (2017)[12] to better disentangle the single effects of the main GHGs on temperature. As the data are given in concentration as parts-per-billion (ppb), we transform them into radiative forcings by using the transformations from Hansen, Sato, Lacis, et al. (1998) as grouped in Table 4.1.

Table 4.1: Radiative Forcings Conversions Formulae

Variable	**Radiative forcing**	**Pre-ind. concentration**
CO_2	$F = f(c) - f(c_0)$ where $f(c) = 5.04 \ln \left[c + 0.0005c^2\right]$	$c_0 \approx 280$ppm
CH_4	$F = 0.04\left(\sqrt{m} - \sqrt{m_0}\right) - [g(m, n_0) - g(m_0, n_0)]$ where $g(m, n) = 0.5 \ln \left[1 + 0.00002(mn)^{0.75}\right]$	$m_0 \approx 700$ppb
N_2O	$F = 0.14(\sqrt{n} - \sqrt{n_0}) - [g(m_0, n) - g(m_0, n_0)]$	$n_0 \approx 275$ppb

We integrate the three time series in place of **G** in the previous model set up, obtaining a total of 10 series spanning the timeframe from 1871 to 2014. Figure 4.11 displays the time series of the three main GHGs.

[12] Data available at `https://www.climatecollege.unimelb.edu.au/cmip6`

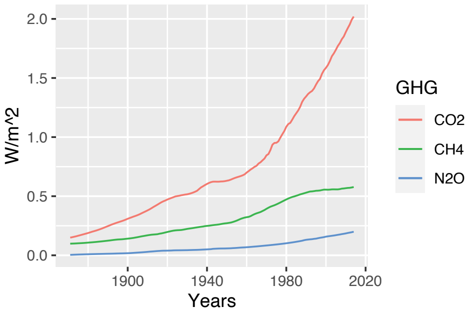


Figure 4.11: Climate Time Series

From Figure 4.12 we do not observe direct causal relations between the three main GHGs and temperature anomalies. However, many indirect relations are observed. In Figure 4.15-4.17 we highlight the amount of possible causal paths between, in turn, CO_2, CH_4, N_2O and temperature anomalies.

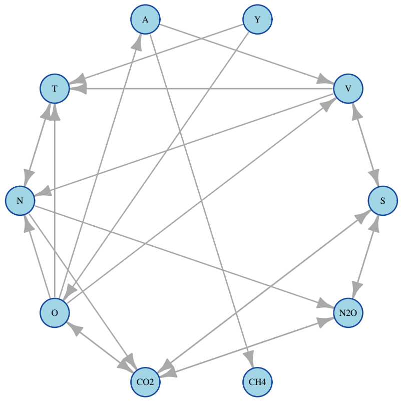


Figure 4.12: Climate Network, $\alpha = 0.1$, $p = 3$

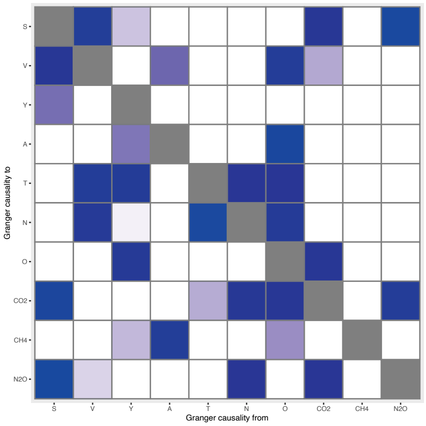
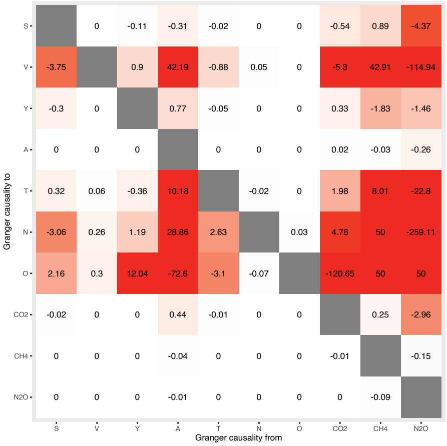


Figure 4.13: P-values heat-map



	S	V	Y	A	T	N	O	CO2	CH4	N2O
S		0	-0.11	-0.31	-0.02	0	0	-0.54	0.89	-4.37
V	-3.75		0.9	42.19	-0.88	0.05	0	-5.3	42.91	-114.94
Y	-0.3	0		0.77	-0.05	0	0	0.33	-1.83	-1.46
A	0	0	0		0	0	0	0.02	-0.03	-0.26
T	0.32	0.06	-0.36	10.18		-0.02	0	1.98	8.01	-22.8
N	-3.06	0.26	1.19	28.86	2.63		0.03	4.78	50	-259.11
O	2.16	0.3	12.04	-72.6	-3.1	-0.07		-120.65	50	50
CO2	-0.02	0	0	0.44	-0.01	0	0		0.25	-2.96
CH4	0	0	0	-0.04	0	0	0	-0.01		-0.15
N2O	0	0	0	-0.01	0	0	0	0	-0.09	

Figure 4.14: Magnitude of Coefficients

We observe a total of 11 paths from CO_2 to T and 17 from N_2O to T. Surprisingly, we do not find paths from CH_4, which in fact results as an end node in the network. One possible explanation would be the masking (cooling) effect of the aerosols. As observed in Hansen, Sato, Kharecha, et al. (2017) "*the elevated GHG concentrations induce a radiative forcing that in turn would cause more than the observed recent global warming if it were not for the cooling effect by aerosols*". In fact, Figure 4.12 shows how tropospheric aerosols & surface albedo leads both methane and stratospheric aerosols. Finally, Figure 4.18 displays the causal paths from GDP to temperature. We observe a total of 15 paths passing through all the variables except for CH_4. This gives clear evidence of how GDP has a broad, outsourcing interplay through time with several climate variable in the span of 3 years. In general, we observe temperature being caused directly from O, V, Y, N as in the analysis in Section 4.4.2.

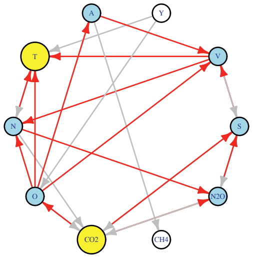


Figure 4.15: CO_2 to T paths (11)

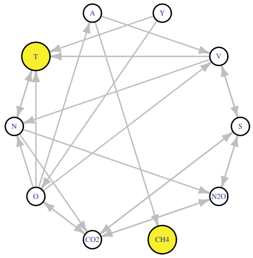


Figure 4.16: CH_4 to T paths (0)

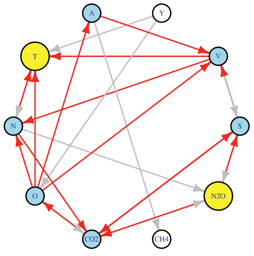


Figure 4.17: N_2O to T paths (17)

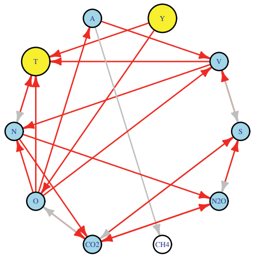


Figure 4.18: Y to T paths (15)

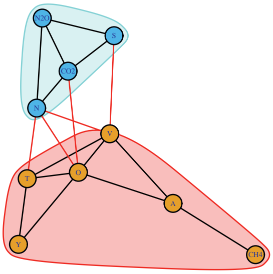


Figure 4.19: Modularity-based Clusters

From Figure 4.12 we observe few *feedback* relations: as observed in Section 4.4.2 a bijective relationship is established between ENSO and temperature. We also observe a feedback between CO_2 and N_2O, one between CO_2 and ocean heat content, one between CO_2 and solar activity, one between N_2O and solar activity and finally one between solar activity and stratospheric aerosols.

Many more loops are now observed compared to Section 4.4.2, well demonstrating the complex, broad interplay among variables in climate systems. Even though no Eulerian cycle[13] is observed, almost all the variables in the system, except Y and CH4, have paths starting and

[13] in Graph Theory an Eulerian cycle is a paths that starts and ends at the same vertex, visiting all other vertices in its route.

ending in their own vertex. We only mention here those cyclic-paths involving temperature and greenhouse gases. For cycles starting respectively from CO_2 and N_2O we find the following cyclic (simple) paths as reported in Figure 4.20, 4.21:

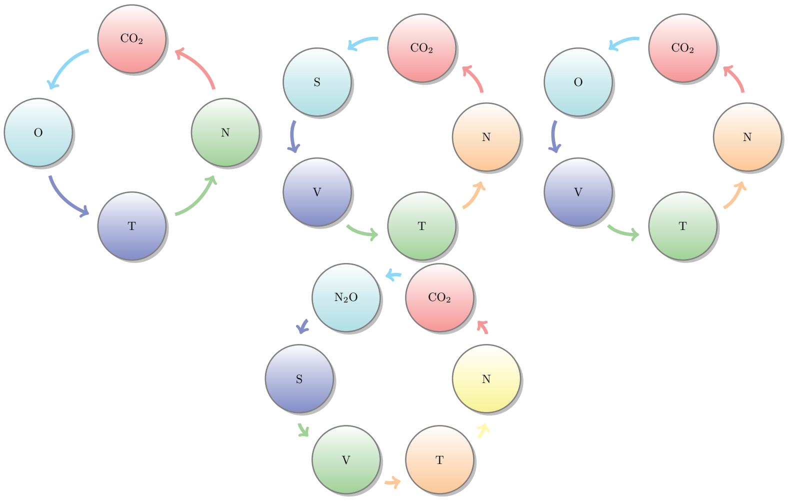


Figure 4.20: Cyclical paths from/to CO_2 via T

What stands out from these cycles effects is that ENSO (N) is a crucial variable to account for in analyzing a climate system. In fact, ENSO appears in all the circular causal chains in Figures 4.20, 4.21 starting from greenhouse gases and passing through temperature. Stratospheric aerosols (V) follow closely as they appear in 8/10 of the reported cycles. Tropospheric aerosols (A) instead only appears in one cycle. Both solar activity (S) and ocean heat content (O) appear in 5/10 of the reported cycles and finally GDP never appears within any cycle.

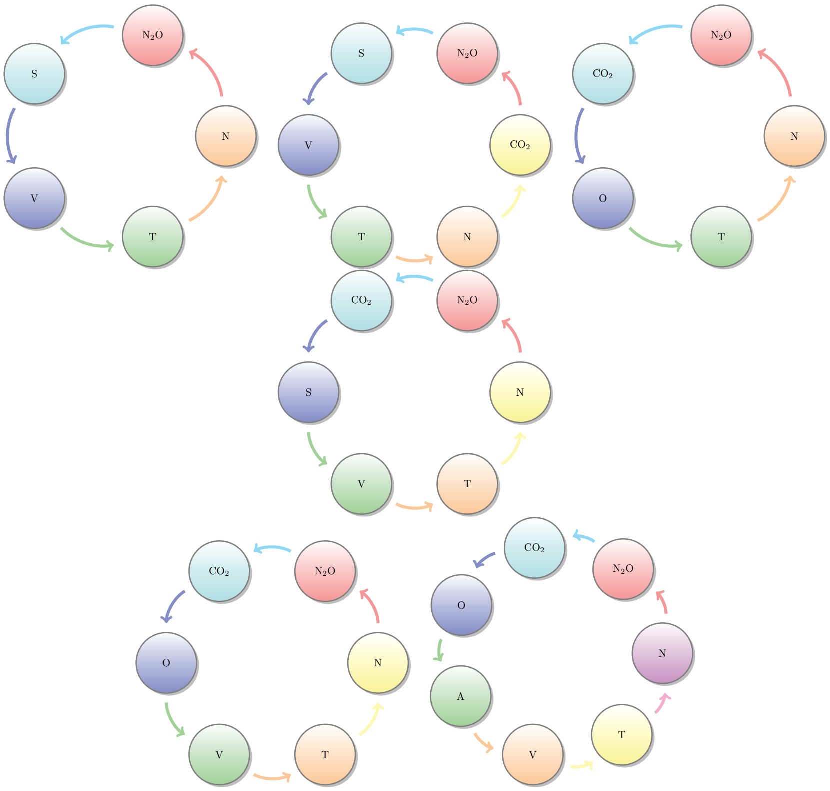


Figure 4.21: Cyclical paths from/to N$_2$O via T

In order to detect communities in the causal network, in Figure 4.19 we employ the hierarchical agglomeration algorithm of Clauset et al. (2004). This employs a greedy optimization strategy in which each vertex of the (undirected) graph is initially considered as unique member of a community of one and repeatedly two communities whose amalgamation produces the largest increase in *modularity* are joined together. Let A_{vw} be the element of the adjacency matrix which constitutes our estimated network i.e., $A_{vw} = 1$ if vertices v and w are connected and zero otherwise. $m = 1/2 \sum_{vw} A_{vw}$ represents the number of edges in the graph and $k_v = \sum_w A_{vw}$ the *degree* of a vertex v, namely the number of edges incident upon such vertex. For a randomized network, $k_v k_w/2m$ would then be the expected fraction of within-community edges. Letting vertices being divided into *communities* where v belongs to community c_v and w belongs to community c_w, then modularity (M) is defined as $M = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m}\right] \delta\left(c_v, c_w\right)$, where $\delta\left(c_v, c_w\right) = 1$ if $i = j$ and $\delta\left(c_v, c_w\right) = 0$ otherwise. In other words, if the fraction of within-community edges in the estimated network equals, or it is close to, the expected one from a randomized network, then modularity will be zero or close to it. Vice-versa modularity will be positive. In Figure 4.19 we find a modularity of $M = 0.208$ and two distinct clusters of community size 6 and 4 respectively are identified. The larger cluster contains both tropospheric and stratospheric aerosols as well as temperature, ocean heat content and, interestingly, GDP and CH_4. The smaller cluster contains CO_2, N_2O, solar activity and El Niño. The connections that go over their community cluster are: solar activity with stratospheric aerosols, CO_2 with ocean heat content and El Niño with both temperature, ocean heat content and stratospheric aerosols. The result of the clustering is in line with the cyclical effects described earlier in Figures 4.20,4.21. In fact, CO_2, N_2O are clustered together with ENSO (N) and solar activity (S) which are crucial variables concerning cyclical effects from greenhouse gases to themselves through temperature.

The outlined methodology in Chapter 3 also allows for testing causality among blocks as explicit in Algorithm 3, rather than just conditional bivariate tests. This implies the possibility of verifying whether a block of variables is Granger causal for another block or alternatively whether a block is Granger causal for a single variable. In our context, we could be interested in testing whether the disaggregated greenhouse gases block-Granger causes temperature directly as we observed in Section 4.4.2. Should be noted that the aggregated series of greenhouse gases in Section 4.4.2 contains also chlorofluorocarbons while in our disaggregated analysis we had to exclude them because of data limitations.

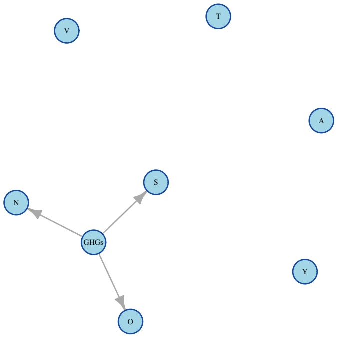


Figure 4.22: Block GHGs connections, $\alpha = 0.1$, $p = 3$

In Figure 4.22 we report only the connections between the block of greenhouse gases i.e., $GHGs = (CO_2, N_2O, CH_4)$ and we observe a total of three connections with ocean heat content, ENSO and solar activity but no direct block-causality with temperature as well as aerosols or production. We also observed the direct connection with ocean heat content in Section 4.4.2 and similarly the connections with solar activity and ENSO, although indirect.

In the following Section 4.5 we are going to perform sensitivity analysis on the lag-length p. One reason why we do not observe direct connections between greenhouse gases and temperature is the slow response that temperature has to changing in greenhouse gases. It might therefore be beneficial to manually enlarge the lag-length considered to verify these relations.

4.5 Sensitivity Analyses

4.5.1 Unit Root Testing

As clearly visible from the various time series plots in Figure 4.1, 4.5,4.11 climate variables are affected by stochastic and/or deterministic trends. Hereby in Table 4.2 we report the p-values for the *autoregressive wild* *bootstrap ADF* test[14] (see Friedrich et al., 2020) calculated on 1999 replications using a significance level of 5%, based on the union of rejections of four tests with different number of deterministic components and different types of detrending (intercept only, intercept plus trend, OLS detrending, quasi-differenced detrending), as in Smeekes and Taylor (2012). We also report the amount of differences δ needed to make the series stationary according to the same test.

Table 4.2: Unit Root Tests

Variable	S	V	Y	G	A	T	N	O	CO_2	CH_4	N_2O
p-value	0.45	0.006	0.43	0.57	0.75	0.97	0.001	0.93	0.65	0.68	0.97
Δ^δ	1	0	1	2	2	1	0	2	2	2	2

Controlling for false discovery rate at level 0.05 within the same set of tests returns the same conclusions as Table 4.2. A Johansen trace test for cointegration (with linear trend) returns 5 cointegrating relations among the series in Table 4.2 excluding G. We observe, as mentioned earlier in Section 4.1, a certain degree of persistence in the series, especially greenhouse gas, both in the aggregated and disaggregated formats, tropospheric aerosols & surface albedo and ocean heat content which require two differences to become stationary. What is remarkable of the methodology outlined in Section 4.3 is that this unit root analysis is completely unnecessary. We do not need to know the integration orders and we do not need to apply transformations to the series. We only need to apply a lag-augmentation d which we believe being sufficient to render the series stationary. As argued earlier, and as verified empirically in Table 4.2, $d = 2$ is always the recommended augmentation.

[14] R code for the autoregressive wild bootstrap ADF union test is available within the package *bootUR*.

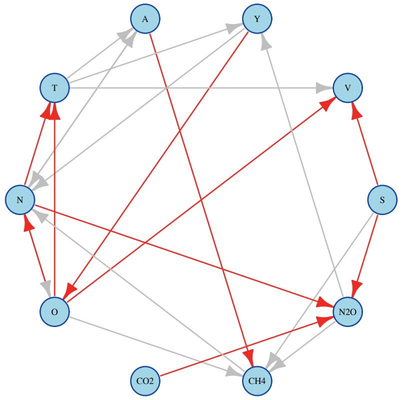


Figure 4.23: Climate Network, $\alpha = 0.1$, $p = 6$

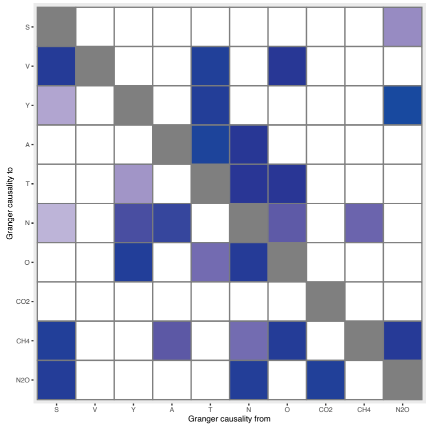


Figure 4.24: P-values heat-map

To investigate the difference with our method, we transform the series to stationary taking the appropriate differences according to Table 4.2. Then we reproduce the disaggregated analysis of Section 4.4.3. To do so, we employ the PDS-LM algorithm of Chapter 2 designed for testing Granger causality in high dimensional stationary VARs. In other words, we employ an entirely analogous algorithm to what we specified in Section 4.3 with the only difference that no lag-augmentation is applied to the system. Figure 4.23, 4.24 display the same network and p-values heat map as for Figure 4.12,4.13 but now with the stationary-transformed series. First, we observe an overall smaller amount of connections (22) in the stationary analysis compared to the non-stationary one (24). In red are highlighted the common connections (10) between network Figure 4.12 and network Figure 4.23. Notably, the estimated lag-length is double than in the non-stationary analysis reflecting the need of a longer lag-length to account for the long run relations. Overall, we observe how several connections present in levels are wiped out when considering the differences, thus avoiding biased unit roots testing is beneficial as to obtaining a richer and more robust picture of the connections in a climate system.

4.5.2 Lag-Length Increases

So far we employed a data-driven selection of the lag-length p. Its estimation is crucial in a VAR context and even more so in this lag-augmented one as the additional lag-augmentation d depends on p (see Section 4.3). Even though the employed methodology is an upper bound, still we could be interested in observing what is the behavior of the disaggragated greenhouse gas network as in Figure 4.12, when the lag-length is manually augmented. This is interesting as on the one hand it shows how our modeling framework is able to handle larger dimensionalities than strictly speaking those we considered here. On the other hand, it is interesting as it is known in the climate science literature how temperature exhibits a slow response to changes in many climate variables, including greenhouse gases. As such, considering larger lag-lengths can open up to new connections masked by the too strict lag-length previously considered. So far we estimated $p = 3$, so the coefficients we reported in Section 4.4.3 were expressing a compound effect of the relevant Granger causing variable to the Granger caused in the span of 3 past years. We now consider the same analysis for a sequence of lag-lengths $p = (10, 15, 30)$. This means that we have up to 302 variables to estimate per equation for a given sample size of 144 data points, thus really exploiting the high-dimensional capabilities of our method.

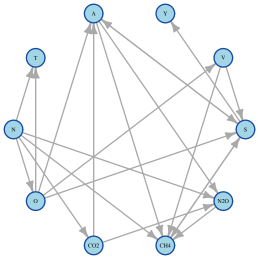


Figure 4.25: $\alpha = 0.1$, $p = 10$

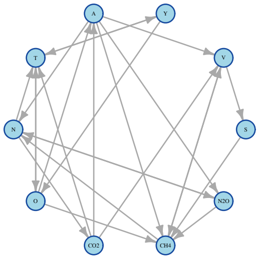


Figure 4.26: $\alpha = 0.1$, $p = 15$

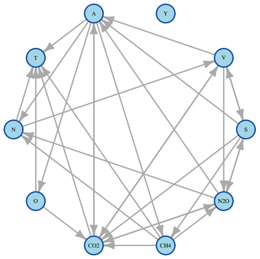


Figure 4.27: $\alpha = 0.1$, $p = 30$

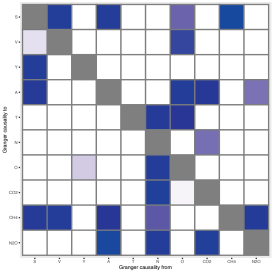


Figure 4.28:
P-values heat-map

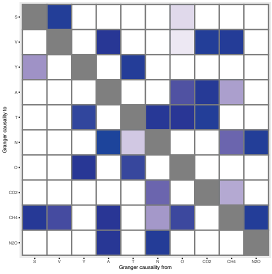


Figure 4.29:
P-values heat-map

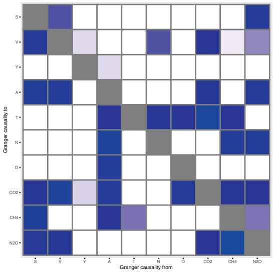


Figure 4.30:
P-values heat-map

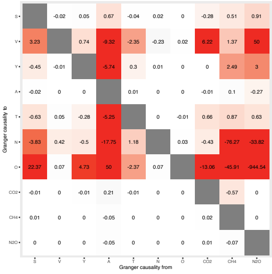


Figure 4.31:
Coeff. heat-map

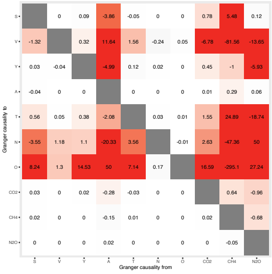


Figure 4.32:
Coeff. heat-map

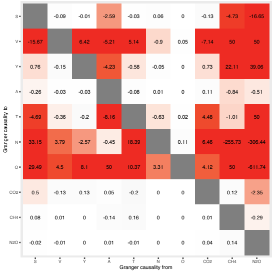


Figure 4.33:
Coeff. heat-map

Figure 4.25, 4.26, 4.27 report the networks estimated respectively with $p = 10, 15, 30$ lags. Similarly, Figure 4.28, 4.29, 4.30 display the respective p-value heat maps and Figure 4.31, 4.32, 4.33 the respective compounded coefficients heat maps. From an overall perspective and comparing the results with Figure 4.12 where we had $p = 3$, we find a comparable total amount of connections in the systems: for $p = 3$ is 24 for $p = 10$ is 21 for $p = 15$ is 24 and for $p = 30$ is 30 connections. Focusing on those connections involving temperature we find that at all lag-lengths considered ENSO and ocean heat content are found to strongly Granger cause temperature. Production is also found to Granger cause temperature but only when $p = 3$ and $p = 15$. As expected, accounting for many more years within the lags is particularly beneficial to uncover direct connections between disaggregated greenhouse gases and temperature. We find at $p = 15$ that CO_2 is Granger causal for temperature and at $p = 30$ that both CO_2 and CH_4 are Granger causal for temperature. We also uncover a direct connection between tropospheric aerosols & surface albedo and temperature at $p = 30$. The dynamic among the greenhouse gases seem to also be enhanced by the larger lag-length. At $p = 30$ we find CO_2 and N_2O having a feedback causal relation and CH_4 to be Granger causal both for CO_2 and N_2O. Especially for CH_4 which in the analysis of Section 4.4.3 resulted as being an end-node, the larger lag-length has uncovered many connections previously not visible. For GDP the result seems more mixed: while at 15 lags it shows an interesting bijective relation with temperature at 30 lags such relation completely disappears leaving GDP unconnected with the rest of the system.

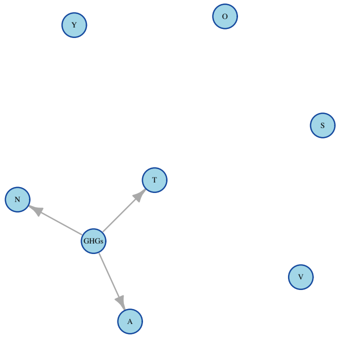


Figure 4.34: Block GHGs connections, $\alpha = 0.1$, $p = 15$

In Figure 4.34 we repeat the block-Granger causality analysis as at the end of Section 4.4.3 using $p = 15$ lags. We now find the greenhouse gases considered i.e., CO_2, N_2O and CH_4, when considered as a block, they do Granger cause temperature as expected and as already observed in Section 4.4.2.

4.6 Discussion and Concluding Remarks

We employ high-dimensional Granger causality tests to investigate the connections within climate systems. We especially focus on the links between radiative forcings and global temperature. Predictive causality in the sense of Granger coupled with a potentially high-dimensional information set robustify the causal findings. We employ and follow the Granger causality framework of Chapter 3 in designing a lag-augmented post-double-selection framework where honest inference in guaranteed and at the same time no care is needed with respect to the unit-root and cointegration properties of the time series at hand. This is of particular appeal with climate data which contains stochastic trends and exhibits long memory. We build a dataset containing a set of the most relevant climate time series coupled with GDP for a time frame spanning 1850-2019. We consider three scenarios of increasing dimensionality: we start with a simple system only containing few fundamental variables usually employed to describe the climate: the solar activity, aerosols and of course temperature and greenhouse gas concentration. We also add GDP to account for economic effects on climate variables and vice-versa. We use the post-double-selection Algorithm 3 on every pair of variables conditional on the remaining ones, thus to obtain a network of Granger causal connections. We then increase the dimensionality accounting for variables not always accounted for in estimating climate networks, namely the ocean heat content and the El Niño southern oscillation index. We find that the inclusion of such variables uncovers several important causal connections in the estimated network. Finally, we consider a disaggregated setting where we decompose the greenhouse gases into their three main components: CO_2, CH_4 and N_2O and we identify indirect causal paths from each of them to temperature, as well as from GDP to temperature. We are also able to apply clustering based on connectedness and as well identify causal feedbacks and causal cycles between greenhouse gases and temperature. As the proposed Algorithm 3 works for conditional blocks-Granger causality and not necessarily only for conditional bivariate tests, we are also able to test whether blocks of variables, as the re-aggregated greenhouse gases, Granger cause temperature. We conclude with a sensitivity analysis in Section 4.5. First, we compare our disaggregated analysis in levels with the same analysis in differences. Namely, we first test the variables for unit root and cointegration and apply the appropriate differencing order to each of them to make them stationary. By replicating the analysis, we find that many connections gets lost when differencing is directly applied to the series. As we argue, unit root and cointegration pre-tests are inducing biases and are preferably avoided. Also, taking differences of the series wipes out their memory and introduce further bias if cointegration is present. Second, even though we use a data-driven method to select the lag-length via information criteria, we also manually enlarge the lag-length using our previous estimated lag-length as lower bound. We find that for climate time series and especially to uncover causal connections among disaggregated greenhouse gases and temperature is often preferable to consider larger lag-lengths.

Directions for further future research are in order. Conditional on the other variables, impulse response functions of temperature to a shock in greenhouse gases, whether significant, are also a causal discovery although not in the same sense as Granger. Considering the same augmented post-double-selection algorithm within the recent local projections framework (see e.g. Jordà, 2005; Plagborg-Møller and Wolf, 2019) would roubustify the causal discoveries and further shed light on the behavior in future time of global temperature to an increase/decrease of greenhouse gases.

We used temperature anomalies as a metric for global warming. Also other metrics have been used in the literature as e.g., the classical global mean surface temperature or Earth's energy imbalance. As it is this imbalance that drives continued warming, the latter metric, which integrates over all climate forcings, gives an indication about where the climate is heading. If availability of the data allows for it, replicating the analyses with other global warming metrics could robustify the findings.

Granger causality is also investigated in Eichler (2013), where he proposes a two steps algorithm. First, one needs to identify all potential direct causal links among the components of a time series where a potential cause is a variable that Granger & Sims causes another variable. Second, one needs to identify the exact nature of a potential causal link, whether it is a true or a spurious causal link. Depending on whether a certain mediating variable C needs to be included or omitted to make variable A non-causal for variable B, we can label the potential cause C as either true or spurious. We did not follow exactly this route here although our analysis in Section 4.4 somewhat followed the same idea: testing (only) Granger causality in a growing system of variables and observe which relations are spurious and which not. Using the idea of impulse responses with local projections, coupled with our Granger causality test would make it for a valid combined procedure that can robustify the causal findings further.

5
Dynamic Factor Models with Sparse VAR Idiosyncratic Components[1]

[1] This chapter is based on a joint work with Jonas Krampe from University of Mannheim.

Abstract

In this chapter we reconcile high-dimensional sparse and dense techniques within the framework of a Dynamic Factor Model and assume the idiosyncratic term follows a sparse vector autoregressive model (VAR). The different diverging behavior of the eigenvalues of the covariance matrix allows to disentangle the two different sources of dependence. The estimation is articulated in two steps: first, the factors and their loadings are estimated via principal component analysis and second, the sparse VAR is estimated by regularized regression on the estimated idiosyncratic components. We prove consistency of the proposed estimation approach as the time and cross-sectional dimension diverge. We complement our procedure with a joint information criteria for the VAR lag-length and the number of factors. The finite sample performance of our procedure is illustrated by means of a thorough simulation exercise.

5.1 Introduction

With the increased availability of large dimensional datasets and the need of techniques able to handle them, the econometrics literature has adapted and rapidly grown in the last years. Datasets containing large amounts of variables (N) with respect to the sample size (T) are loaded with information and although this represents a great potential to be exploited (the *blessings*), it also carries not few troubles for the statistician to deal with (the *curses*). As the parameter space expands at fast speed, its elements to estimate soon start to be too many for the sample information available to reliably estimate them. Overfitting and surging variance indeed cause failure of standard methods designed for settings where N is small relative to T. In one way or another, the core idea to get away from the curses is: *dimensionality reduction.* The econometric literature mostly polarizes on either factor models or penalized regression techniques. The former assumes the behavior of an economic variable is sensibly decomposed into a component driven by few unobservable (latent) factors, which are common to many other economic variables but load differently on each of them, and a variable specific idiosyncratic component. The dimensionality reduction obtained via the estimation of the factors is such that the explained variability of the original set of variables is mostly preserved, hence the *dense* label. On the other hand, penalized regression techniques yields dimensionality reduction working with a *sparsity* assumption over the underlying true model. Namely: sparsity limits the number of direct channels to affect other variables. Both factor models and sparse-regression techniques are widely employed in practice. However, usually of the two, one excludes the other. The reason being the radically different approach, namely *dense* versus *sparse.* Neither of the two school of thoughts are free of criticisms: regularization techniques because of the sparsity assumption, often seen as a too strong assumption especially in macroeconomics; factor models because they need some consistent criterion to select a certain -not too large- amount of factors. Many papers have compared empirically the performances of either of the approaches, especially in terms of macroeconomic forecasting, among others: Smeekes and Wijler (2018), Coulombe, Leroux, et al. (2020), and Medeiros, Vasconcelos, et al. (2021) or criticized each other approaches (Giannone, Lenza, et al., 2017; Fava and Lopes, 2020).

In this Chapter we reconcile the two worlds of dense and sparse modeling by means of focusing on exploiting the positive aspects of both. We work within the framework of a Dynamic Factor Model where we employ principal component analysis (PCA) to estimate the factors and high-dimensional penalized vector autoregressive model (VAR) through the adaptive lasso in order to estimate the idiosyncratic components. This approach is beneficial since it allows to disentangle in the system covariance matrix, the dependence among its diverging eigenvalues (i.e., the factors) with the dependence among the bounded ones (i.e., the idiosyncratic components). Coincidentally, we allow for cross-sectional and time dependence in the idiosyncratic term and we assume this follows a high-dimensional VAR model. We work under the physical (functional) dependence framework of Wu (2005) and we show consistent estimation of both idiosyncratic components and the factors as both the cross-sectional and time dimensions grow large. We also propose a joint information criteria which combines the Bai and Ng (2002), Alessi et al. (2010) approaches with an extra penalty allowing for simultaneous lag-length estimation.

The economic interpretation of factor models is a crucial reason of their wide use. For instance, the literature on Dynamic Stochastic General Equilibrium (DSGE) models and real business cycle (see e.g. Sargent and Sims, 1977) assumes few common forces to drive the whole economy. Factor models were originally envisioned for cross-sectional data and their time-series extension, broadly referred to as *Dynamic Factor Models* (DFM), was first proposed by Geweke (1977). DFMs are nowadays ubiquitous in economics, their applications range over from: macroeconomics forecasting (see among others: Stock and Watson, 1999; Stock and Watson, 2002a; Stock and Watson, 2002b; Forni, Hallin, Lippi, and Reichlin, 2003; Boivin and Ng, 2005; Koopman and Wel, 2013; Marcellino et al., 2016; Forni, Giovannelli, et al., 2018), real-time monitoring (*nowcasting*) (see among others: Giannone, Reichlin, et al., 2008; Aruoba, Diebold, et al., 2010; Aruoba and Diebold, 2010; Schiavoni et al., 2021), international business cycle (see among others: Kim and Nelson, 1998; Lee, 2012; Lee, 2013; Doz et al., 2020), construction of leading indicators (see among others: Stock and Watson, 1989b; Altissimo et al., 2001; Forni, Hallin, Lippi, and Reichlin, 2001; Banerjee et al., 2005), to monetary policy applications (see among others: Bernanke et al., 2005; Forni and Gambetti, 2010; Korobilis, 2013).

Variable selection procedures aimed at selecting and estimating only the subset of truly relevant variables have been introduced via different ℓ_q-norm penalizations of the least-square minimization problem. Ridge regression (Hoerl and Kennard, 1970), the lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), adaptive lasso (Zou, 2006), scad (Fan and Li, 2001) are only few among the plethora of techniques developed to tradeoff variance with bias and thus get away from the curse of dimensionality. Many more refinements of such procedures have acquired a great deal of space and interest in the statistics and econometrics literatures. Econometrics applications of these procedures see among others: instrumental variables estimation (see among others: Belloni, Chernozhukov, and Hansen, 2011b; Belloni, Chen, et al., 2012; Belloni, Chernozhukov, and Wang, 2014; Windmeijer et al., 2019), treatment effect models (see among others: Belloni, Chernozhukov, Fernández-Val, et al., 2015; Li and Bell, 2017; Ju et al., 2020), time series models (see among others: Kock and Callot, 2015; Medeiros and Mendes, 2016b).

There already exists applications in the literature combining dynamic factor models with sparse vector autoregressive models, see e.g.,Barigozzi and Hallin (2017) and Barigozzi and Brownlees (2019). However, Barigozzi and Hallin (2017) and Barigozzi and Brownlees (2019) do not present any theoretical results about the combined approach. Also Kneip and Sarda (2011) and Fan, Ke, et al. (2020) combine factors with regularized models. Since regularized methods as the lasso have difficulties with strongly correlated regressors, especially in the context of model selection, their aim is to decorrelate the regressors by adjusting for the factors. Fan, Ke, et al. (2020) extend and generalize the results of Kneip and Sarda (2011) arguing that the combined approach is flexible in the sense that it performs well regarding both prediction and model selection and in both uncorrelated and highly correlated cases. Furthermore, Fan, Masini, et al. (2021) provide hypothesis tests to test whether after removing factors (as well as trends in a first step) the regressors possesses some pre-defined weakly correlated structure or not. Fan, Ke, et al. (2020) and Fan, Masini, et al. (2021) allow for time-dependent regressors, however they do not consider that the idiosyncratic part follows a sparse vector autoregressive model. Moreover, their assumption about the dependency of the idiosyncratic part excludes sparse vector autoregressive models where the cross-sectional sparsity can grow with the sample size. In the context of vector autoregressive processes for network modeling of high-dimensional time series data, a similar approach is to combine the estimation of a low-rank matrix and a sparse matrix as it is done in Basu, Li, et al. (2019). The low-rank part takes here a similar role as the common component of DFMs while sparsity is assumed over the autoregressive polynomial matrix of the idiosyncratic term. The combination of low-rank plus sparse has been also explored in the context of an approximate factor model by Lin and Michailidis (2019). However, the approaches in Basu, Li, et al. (2019) and Lin and Michailidis (2019) differ from ours of combining DFM and sparse VARs. First note that these two approaches do not describe the same model. Furthermore, the sparsity in Basu, Li, et al. (2019) and Lin and Michailidis (2019) is far more restrictive than the sparsity considered here for the idiosyncratic component and their results are useful in quantifying the overall estimation and prediction error whereas in this Chapter also the prediction error for a single time series can be quantified. A more detailed discussion can be found in Section 5.3.2.

The remainder of the Chapter is organized as follows: Section 5.2 introduces the dynamic factor model with sparse VAR idiosyncratic components and report few standard assumptions defining its behavior. Section 5.3 is devoted to describe the two-step procedure used to estimate the DFM with sparse VAR idiosyncratic components and prove its consistency. Theorem 1 derives a representation of the idiosyncratic components estimation error while Theorem 2 is the main result establishing bounds for the estimation error for the second step of the estimation procedure, i.e., for the lasso on the sample estimates of the idiosyncratic component. Section 5.4 considers the problems of: estimating the number of factors, determining the lag-length in the VAR and tune the penalty parameter for the lasso. Section 5.5 report simulation results for our proposed method under different VAR data generating processes in terms of design and sparsity. Finally Section 5.6 concludes.

A few words on notation. Throughout the Chapter we use boldface characters to indicate vectors and boldface capital characters for matrices. For any n-dimensional vector $\boldsymbol{x}$, we let $\|\boldsymbol{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ denote the ℓ_p-norm and $\boldsymbol{e}_j = (0, \dots, 0, 1, 0, \dots, 0)^\top$ denotes a unit vector of appropriate dimension with the one appearing in the jth position. Furthermore, for a $r \times s$ matrix $\boldsymbol{A} = (a_{i,j})_{i=1,\dots,r,j=1,\dots,s}$, $\|\boldsymbol{A}\|_1 = \max_{1\le j\le s} \sum_{i=1}^r |a_{i,j}| = \max_j \|\boldsymbol{A}\boldsymbol{e}_j\|_1$, $\|\boldsymbol{A}\|_\infty = \max_{1\le i\le r} \sum_{j=1}^s |a_{i,j}| = \max_i \|\boldsymbol{e}_i^\top \boldsymbol{A}\|_1$ and $\|\boldsymbol{A}\|_{\max} = \max_{i,j} |\boldsymbol{e}_i^\top \boldsymbol{A}\boldsymbol{e}_j|$. Denote the largest absolute eigenvalue of a square matrix $\boldsymbol{A}$ by $\rho(\boldsymbol{A})$ and let $\|\boldsymbol{A}\|_2^2 = \rho(\boldsymbol{A}\boldsymbol{A}^\top)$. We denote the smallest eigenvalue of a matrix $\boldsymbol{A}$ by $\sigma_{\min}(A)$. For any index set $S \subseteq \{1, \dots, n\}$, let $\boldsymbol{x}_S$ denote the sub-vector of $\boldsymbol{x}_t$ containing only those elements x_i such that $i \in S$. $\|\boldsymbol{x}\|_0$ denotes the number of non-zero elements of $\boldsymbol{x}$.

5.2 The Model and Notation

Let $\boldsymbol{X} = (x_{1,t}, \dots, x_{N,t})$ for $t = 1, \dots T$ be a $N \times T$ rectangular data array representing a finite realization of an underlying real-valued stochastic process $\{x_{i,t}\}$. Assume $x_{i,t}$ can be decomposed into the sum of a common component $\chi_{i,t}$ and an idiosyncratic component $\xi_{i,t}$, both unobservable. The common component $\chi_{i,t}$ is driven by an r-dimensional vector of common factors $\boldsymbol{f}_t = (f_{1,t}, \dots, f_{r,t})^\top$ where $r \perp\!\!\!\perp N$, $r \ll N$ and each of which has a certain specific loading $\ell_{i,t}$. We consider the number of factors r to be fixed as both the cross sectional dimension N and the time series dimension T grow large. This is a reasonable claim: assuming r to be a strictly increasing function in N or T would be tantamount to assume that all the eigenvalues of a large dimensional covariance matrix would necessarily diverge as the dimensions increase which would clearly not be reasonable. Nonetheless, r needs to be estimated from the data and we consider this issue jointly with the lag-length estimation later in Section 5.4. We do not assume the common components $\boldsymbol{f}_t$ to be independent and identically distributed but we allow them instead to be dynamic. However, the dynamic does not directly links with $\boldsymbol{x}_t$, thus making the relationship between $\boldsymbol{x}_t$ and $\boldsymbol{f}_t$ still static. This differs from the framework of Forni, Hallin, Lippi, and Reichlin (2000) which assumes a pervasive dynamic of the common factors where $\boldsymbol{x}_t$ is set to also depend on $\boldsymbol{f}_t$ with lags in time. In other words, in this Chapter we are working with the *Dynamic Factor Model* where both factors and idiosyncratic components are allowed to be stationary stochastic processes. Let the common component vector $\boldsymbol{\chi}_t = (\chi_{1,t} \cdots \chi_{N,t})^\top$ and the idiosyncratic component vector $\boldsymbol{\xi}_t = (\xi_{1,t} \cdots \xi_{N,t})^\top$, then the factor model decomposition takes the following usual form

$$\boldsymbol{x}_t = \boldsymbol{\chi}_t + \boldsymbol{\xi}_t. \tag{5.1}$$

Now, the common components $\chi_{i,t}$ can be expressed as the following linear combination:

$$\chi_{i,t} = \ell_{i,1} f_{1,t} + \ell_{i,2} f_{2,t} + \cdots + \ell_{i,r} f_{r,t} = \boldsymbol{\Lambda}_i^\top \boldsymbol{f}_t. \tag{5.2}$$

Note that $\chi_{i,t}$ is uniquely defined. But since for some rotation matrix $\boldsymbol{H}$, $\chi_{i,t} = \boldsymbol{\Lambda}_i^\top \boldsymbol{H} \boldsymbol{H}^{-1} \boldsymbol{f}_t$ is a valid linear combination as well, $\boldsymbol{\Lambda}_i^\top, \boldsymbol{f}_t$ are only identified up to some arbitrary rotation. Additionally, assume the idiosyncratic component $\boldsymbol{\xi}_t$ to follow a sparse vector autoregressive (VAR) model of order p, where p is the lag-length, as

$$\boldsymbol{\xi}_t = \sum_{j=1}^{p} \boldsymbol{A}^{(j)} \boldsymbol{\xi}_{t-j} + \boldsymbol{v}_t = \sum_{j=0}^{\infty} \boldsymbol{B}^{(j)} \boldsymbol{v}_{t-j} \tag{5.3}$$

for $\boldsymbol{v}_t$ being a zero-mean noise process described in Assumption 10 below. By estimating the factors $\boldsymbol{f}_t$ through standard Principal Components Analysis (PCA) and the sparse VAR models of the idiosyncratic components $\boldsymbol{\xi}_t$ via sparse penalized regression techniques we combine respectively a dense modeling approach with a sparse one and we are able to better capture and disentangle both the dependence among diverging eigenvalues of $\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)$, i.e., the factors, as well as the dependence among the non-diverging eigenvalues of $\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)$, i.e., the idiosyncratic components. To aid the intuition of a VAR modeling of the idiosyncratic component, one can think for instance, at the asset pricing models, designed to explain asset returns through several factors of risk. The common components would represent here the systematic, unobserved, part of information explaining the asset return, in other words: those risk components which, systemically interconnected, decide the level of the asset return. The idiosyncratic part instead, is the whole remaining non-systematic or individual part of information pertaining to the single assets which also contributes in deciding the return of the asset. It follow that in order to model the linear dependence among the idiosyncratic part, the most reasonable choice falls for VARs. To proceed with our theoretical results, we first state some assumptions. In the following Assumptions 9, 10, and 11, the sparsity and stability conditions, the factors, moment conditions, and loadings are further specified.

Assumption 9. (*Sparsity and stability*)
(*i*) Let $\mathfrak{A}$ denote the stacked (companion) VAR matrix of (5.3). Let k denote the row-wise sparsity of $\mathfrak{A}$ with approximate sparsity parameter $q \in [0, 1)$, i.e.,

$$\max_i \sum_{k=1}^{p} \sum_{j=1}^{N} |\boldsymbol{A}_{i,j}^{(k)}|^q = \max_i \sum_{j=1}^{Np} |\mathfrak{A}_{i,j}|^q \leq k.$$

(*ii*) The VAR processes is considered as stable such that for a constant $\rho \in (0, 1)$ we have independently of the sample size T: $\|\mathfrak{A}^j\|_2 = \sqrt{\lambda_{\max}(\mathfrak{A}^{j\top}\mathfrak{A}^j)} \leq M\rho^j$, where M is some constant. Additionally, we have $\|\Gamma_{\xi(0)}\|_\infty \leq k_\xi M$, where $\Gamma_\xi(0) = \mathrm{Var}(\boldsymbol{\xi}_t)$ and $\sigma_{\min}(\mathrm{Var}((\boldsymbol{\xi}_t^\top, \ldots, \boldsymbol{\xi}_{t-p+1}^\top)^\top)) > \alpha$. The sparsity parameter k as well as k_ξ can grow with the sample size.

Assumption 10. (*Factor dynamic and moments*)
The factors are given by a one-sided linear filter with geometrically decaying coefficients, that is:

$$\boldsymbol{f}_t = \sum_{j=0}^{\infty} D^{(j)} \boldsymbol{u}_{t-j},$$

and $\|D^{(j)}\|_2 \leq K\tilde{\rho}^j$, where K is some positive constant and $\tilde{\rho} \in (0,1)$. Furthermore, $\{(\boldsymbol{u}_t^\top, \boldsymbol{v}_t^\top)^\top, t \in \mathbb{Z}\}$ is an i.i.d. sequence and $\mathrm{Cov}(\boldsymbol{u}_t, \boldsymbol{v}_t) = 0$. Let $\zeta > 8$ be the number of finite moments of $\{(\boldsymbol{u}_t^\top, \boldsymbol{v}_t^\top)^\top, t \in \mathbb{Z}\}$, i.e., $\mathbb{E}|\boldsymbol{u}_{t,j}|^\zeta \leq M$ and $\max_{\|\boldsymbol{w}\|_2 \leq 1} \mathbb{E}|\boldsymbol{w}^\top \boldsymbol{v}_t|^\zeta \leq M$. We denote $\boldsymbol{\Sigma}_u =: \mathrm{Var}(\boldsymbol{u}_t)$ and $\boldsymbol{\Sigma}_v =: \mathrm{Var}(\boldsymbol{v}_t)$.

Assumption 11. (*Factors and loadings*)
Let M be some finite constant, then

1. $\lim_{T\to\infty} 1/T \sum_{t=1}^T \boldsymbol{f}_t \boldsymbol{f}_t^\top = \mathbb{E}[\boldsymbol{f}_t \boldsymbol{f}_t^\top] = \boldsymbol{\Sigma}_F \in \mathbb{R}^{r\times r}$ positive definite and $\|\boldsymbol{\Sigma}_F\|_2 \leq M$.
2. $\lim_{N\to\infty} 1/N \sum_{i=1}^N \boldsymbol{\Lambda}_i \boldsymbol{\Lambda}_i^\top = \boldsymbol{\Sigma}_\Lambda$, $\boldsymbol{\Sigma}_\Lambda$ positive definite, $\|1/N \sum_{i=1}^N \boldsymbol{\Lambda}_i \boldsymbol{\Lambda}_i^\top\|_2 \leq M$ for all N, $\|\boldsymbol{\Lambda}\|_{\max} \leq M$ and $\|\boldsymbol{\Sigma}_\Lambda\|_2 \leq M$.
3. All eigenvalues of $\boldsymbol{\Sigma}_F, \boldsymbol{\Sigma}_\Lambda$ are distinct.

Some comments on the above assumptions are in order. Assumption 9 and 10 imply that $\{\boldsymbol{\xi}_t\}$ is stationary and let the autocovariance function be given by $\boldsymbol{\Gamma}_\xi(s-t) = \mathrm{Cov}(\boldsymbol{\xi}_s, \boldsymbol{\xi}_t)$. Furthermore, Assumption 10 implies that also the factors are a stationary processes such that $\{\boldsymbol{x}_t\}$ itself is indeed stationary. In order to quantify the dependence of the occurring stochastic processes, we use the concept of functional dependence, see Wu (2005). Since this is only necessary for the proofs, we do not introduce the notation here and refer to Remark 5.1 in the appendix. The assumption of a stable and row-wise sparse VAR model is a rather standard assumption in the literature of sparse VAR models, see among others Kock and Callot (2015), Han et al. (2015), and Masini et al. (2019) and for a discussion of different sparsity concepts for VAR models see Krampe and Paparoditis (2021). Here, the weaker assumption of approximate sparsity instead of exact sparsity, i.e., $q = 0$, is used. Note that we only require row-wise sparsity, i.e., a time series in $\boldsymbol{x}_t$ can be directly influenced only by k other time series including their lags. We do not require *column-wise sparsity*, i.e., a single time series in $\boldsymbol{x}_t$ and all its past values can have more than k direct channels to affect the elements of $\boldsymbol{x}_t$. Furthermore, the moment condition in Assumption 10 refers to the situation in which only a finite number of moments, here ζ, are finite. Hence, we do not assume sub-Gaussian processes or similar which is often assumed for sparse VAR processes, see Basu and Michailidis (2015b), Kock and Callot (2015), and Han et al. (2015). For sub-Gaussian processes and in the error bounds obtained later on, the polynomial terms depending on ζ would vanish which results in sharper error bounds. The reason for only assuming ζ finite moments is to be more in line with the classical factor literature, see among others Bai (2003), Stock and Watson (2002a), Forni, Hallin, Lippi, and Reichlin (2000), and Forni, Hallin, Lippi, and Zaffaroni (2017). E.g., Bai (2003) derived his inferential results for factor models under 8th finite moments of the idiosyncratic part and 4th finite moments of the factors. Note that the filter in Assumption 10 can be the one-sided representation of a rational filter as in Assumption 2 in Forni, Hallin, Lippi, and Zaffaroni (2017), that is $b(L) = c(L)/d(L)$, where d has no roots in the disc $\{c \in \mathbb{C} | |c| \leq \phi\}, \phi > 1$. Assumption 11 is a rather standard assumption in the factor literature, see Stock and Watson (2002a) and Bai (2003). It implies that each of the factors provides a non-negligible contribution to the variance of each component of $\{\boldsymbol{x}_t\}$. We like to point out here that the time and cross-sectional dependence of the idiosyncratic component is only limited by assuming that it follows a sparse VAR model. Furthermore, it is not clear if assuming a sparse VAR model for the idiosyncratic part is a special case of the assumptions to time and cross-section dependence in the factor literature, see among others Bai, 2003, Assumption C or Forni, Hallin, Lippi, and Zaffaroni, 2017, Assumption 4. The reason for this is that the sparsity is not fixed but it can grow with sample size. Nevertheless, the error bounds obtained later on require that the sparsity cannot grow too fast with increasing dimension.

5.3 Estimation

In this section we outline a two-step approach to estimate a DFM with sparse VAR idiosyncratic components and we prove its consistency. For this, let $\boldsymbol{x}_t$, $t = 1, \ldots, T$ be some observations and let $\boldsymbol{X} = \boldsymbol{\chi} + \boldsymbol{\Xi}$ denote the $T \times N$ matrix form of (5.1). Furthermore, $\boldsymbol{\Lambda}$ denotes the $N \times r$ matrix of loadings and $\boldsymbol{F}$ denotes the $T \times r$ matrix of factors such that $\boldsymbol{\chi} = \boldsymbol{F}\boldsymbol{\Lambda}^\top$ is the matrix counterpart of (5.2). Then, an estimation of the factor decomposition can be obtained by using PCA, see among others Bai (2003) and Bai and Ng (2020). For this, let $\boldsymbol{X}/\sqrt{NT} = \boldsymbol{U}_{NT}\boldsymbol{D}_{NT}\boldsymbol{V}_{NT}^\top$ denote a singular value decomposition of $\boldsymbol{X}/\sqrt{NT}$ such that $\boldsymbol{D}_{NT}$ is a diagonal matrix with the singular values arranged in descending order on its diagonal. $\boldsymbol{U}_{NT}$ and $\boldsymbol{V}_{NT}$ are the corresponding left and right singular vectors, respectively. This can further be written as $\boldsymbol{U}_{NT}\boldsymbol{D}_{NT}\boldsymbol{V}_{NT}^\top = \boldsymbol{U}_{NT,r}\boldsymbol{D}_{NT,r}\boldsymbol{V}_{NT,r}^\top + \boldsymbol{U}_{NT,N-r}\boldsymbol{D}_{NT,N-r}\boldsymbol{V}_{NT,N-r}^\top$, where $\boldsymbol{D}_{NT,r}$ is a diagonal matrix with the first r largest singular values, $d_{NT,1}, \ldots, d_{NT,r}$, arranged in descending order on its diagonal, $\boldsymbol{D}_{NT,N-r}$ is a diagonal matrix with the remaining $N-r$ largest singular values, and $\boldsymbol{U}_{NT,r}, \boldsymbol{U}_{NT,N-r}, \boldsymbol{V}_{NT,r}, \boldsymbol{V}_{NT,N-r}$ are the corresponding left and right singular vectors. The estimators of some rotated version of $\boldsymbol{F}$ and $\boldsymbol{\Lambda}$ are then given by $\hat{\boldsymbol{F}} = \sqrt{T}U_{NT,r}$ and $\hat{\boldsymbol{\Lambda}} = \sqrt{N}\boldsymbol{V}_{NT,r}\boldsymbol{D}_{NT,r}$ such that $\hat{\boldsymbol{\chi}} = \hat{\boldsymbol{F}}\hat{\boldsymbol{\Lambda}}^\top$ and $\hat{\boldsymbol{\xi}} = \boldsymbol{x}_t - \hat{\boldsymbol{\chi}}$. We use here the normalization $\hat{\boldsymbol{F}}^\top\hat{\boldsymbol{F}}/T = \boldsymbol{I}_r$ and $\hat{\boldsymbol{\Lambda}}^\top\hat{\boldsymbol{\Lambda}}$ is a diagonal matrix.

We are assuming $\{\boldsymbol{\xi}_t\}$ follows a sparse vector autoregressive model which can be estimated by regularized methods such as the (adaptive) lasso. This idea leads to the following two-step estimation procedure:

1. Perform a singular value decomposition of $\boldsymbol{X}/\sqrt{NT} = \boldsymbol{U}_{NT,r}\boldsymbol{D}_{NT,r}\boldsymbol{V}_{NT,r}^\top + \boldsymbol{U}_{NT,N-r}\boldsymbol{D}_{NT,N-r}\boldsymbol{V}_{NT,N-r}^\top$, where $\boldsymbol{U}_{NT,r}\boldsymbol{D}_{NT,r}\boldsymbol{V}_{NT,r}^\top$ corresponds to the first r singular values. Set $\hat{\boldsymbol{F}} = \sqrt{T}U_{NT,r}$ and $\hat{\boldsymbol{\Lambda}} = \sqrt{N}\boldsymbol{V}_{NT,r}\boldsymbol{D}_{NT,r}$, and $\hat{\boldsymbol{\xi}} = \boldsymbol{x}_t - \hat{\boldsymbol{F}}\hat{\boldsymbol{\Lambda}}^\top$.

2. Let $\hat{\boldsymbol{\xi}}_t^v = (\hat{\boldsymbol{\xi}}_t^\top, \ldots, \hat{\boldsymbol{\xi}}_{t-p}^\top)^\top$. Then, an adaptive lasso estimator for $\boldsymbol{\beta}^{(j)}$, i.e., the jth row of $(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(p)})$, is given by

$$\hat{\boldsymbol{\beta}}^{(j)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{np}} \frac{1}{T-p}\sum_{t=p+1}^{T}(\hat{\xi}_{j,t} - \boldsymbol{\beta}^\top\hat{\boldsymbol{\xi}}_{t-1}^v)^2 + \lambda\sum_{i=1}^{N}|g_i\beta_i|, \quad (5.4)$$

for $j = 1, \ldots, N$ and where λ is a non-negative tuning parameter which determines the strength of the penalty, $g_i, i = 1, \ldots, N$, are weights, for instance, $g_i = 1$ leads to the standard lasso. Let $(\hat{\boldsymbol{A}}^{(1)}, \ldots, \hat{\boldsymbol{A}}^{(p)})$ be matrices corresponding to stacking $\hat{\boldsymbol{\beta}}^{(j)}, j = 1, \ldots, N$.

For a sparse stationary VAR model, deviation bounds and restricted eigenvalue conditions can be established, see Basu and Michailidis (2015b) and Kock and Callot (2015). Given these, the consistency of the lasso can be derived easily. However, as the idiosyncratic component $\{\boldsymbol{\xi}_t\}$ is not observed in our setting and hence needs to be estimated, the regression in Step 2 is performed only with the estimated idiosyncratic component. Hence, the results of Basu and Michailidis (2015b) and Kock and Callot (2015) cannot be applied here. Before analyzing the second step, we in fact need to quantify the estimation error $\boldsymbol{w}_t := \hat{\boldsymbol{\xi}}_t - \boldsymbol{\xi}_t$ coming from the first step. The aim is to quantify the estimation error $\boldsymbol{w}_t$ in quantities like $\|1/T\sum_{t=1}^T(\boldsymbol{\xi}_t + \boldsymbol{w}_t)(\boldsymbol{\xi}_t + \boldsymbol{w}_t)^\top\|_{\max}$. If we simply apply the rate derived in the literature for approximate factor models, see among others Stock and Watson (2002a) and Bai (2003) which derive $\boldsymbol{w}_t = O_P(\max(1/\sqrt{T}, 1/\sqrt{N}))$, we would obtain $\|1/T\sum_{t=1}^T(\boldsymbol{\xi}_t + \boldsymbol{w}_t)(\boldsymbol{\xi}_t + \boldsymbol{w}_t)^\top\|_{\max} = \|1/T\sum_{t=1}^T\boldsymbol{\xi}_t\boldsymbol{\xi}_t^\top\|_{\max} + O_P(\max(1/\sqrt{T}, 1/\sqrt{N}))$. However, this can be improved if we analyze the estimation error $\boldsymbol{w}_t$ more closely. For this, we follow the idea of the decomposition in eq. (6) in Bai and Ng (2020). To elaborate, we have $1/(NT)\boldsymbol{X}\boldsymbol{X}^\top\hat{\boldsymbol{F}} =$

$\hat{\boldsymbol{F}}\boldsymbol{D}^2_{NT,r}$. Plugging in (5.1), and using the rotation matrix

$$\boldsymbol{H}_{NT}^{\top} = (\boldsymbol{\Lambda}^{\top}\boldsymbol{\Lambda}/N)(\boldsymbol{F}^{\top}\hat{\boldsymbol{F}}/T)\boldsymbol{D}^{-2}_{NT,r}, \tag{5.5}$$

we obtain the following representation for the error between the estimated factors and a rotated version of the true factors

$$\begin{aligned}&\hat{\boldsymbol{f}}_t - \boldsymbol{H}_{NT}\boldsymbol{f}_t =\\ &= \frac{1}{NT}\Big[\sum_{i=1}^{N}\sum_{s=1}^{T}\boldsymbol{f}_{\boldsymbol{t}}^{\top}\boldsymbol{\Lambda}_i\xi_{i,s}\hat{\boldsymbol{f}}_s + \sum_{i=1}^{N}\sum_{s=1}^{T}\xi_{i,t}\boldsymbol{\Lambda}_i\boldsymbol{f}_s^{\top}\hat{\boldsymbol{f}}_s + \sum_{i=1}^{N}\sum_{s=1}^{T}\xi_{i,t}\xi_{i,s}\hat{\boldsymbol{f}}_s\Big]\boldsymbol{D}^{-2}_{NT,r}.\end{aligned} \tag{5.6}$$

Similarly, we obtain by symmetry for the loadings

$$\begin{aligned}&(\boldsymbol{H}_{NT}^{\top})^{-1}\boldsymbol{\Lambda}_i - \hat{\boldsymbol{\Lambda}}_i =\\ &= \frac{1}{T}\Big[\sum_{s=1}^{T}\boldsymbol{H}_{NT}\boldsymbol{f}_s\xi_{i,s} + \sum_{s=1}^{T}(\hat{\boldsymbol{f}}_s - \boldsymbol{H}_{NT}\boldsymbol{f}_s)\xi_{i,s}\\ &+ \sum_{s=1}^{T}\boldsymbol{H}_{NT}\boldsymbol{f}_s[\hat{\boldsymbol{f}}_s - \boldsymbol{H}_{NT}\boldsymbol{f}_s]^{\top}(\boldsymbol{H}_{NT}^{T})^{-1}\boldsymbol{\Lambda}_i\\ &\sum_{s=1}^{T}[\hat{\boldsymbol{f}}_s - \boldsymbol{H}_{NT}\boldsymbol{f}_s][\hat{\boldsymbol{f}}_s - \boldsymbol{H}_{NT}\boldsymbol{f}_s]^{\top}(\boldsymbol{H}_{NT}^{T})^{-1}\boldsymbol{\Lambda}_i\Big].\end{aligned}$$

These representations can be used to derive the order of the estimation error for the factors and loadings as it is done with a slightly different rotation matrix in Bai (2003). However, as our focus is not only on the factors and loadings but also on $\boldsymbol{w}_t := \hat{\boldsymbol{\xi}}_t - \boldsymbol{\xi}_t$, we use these results to derive a simpler representation of $\boldsymbol{w}_t$, see the following Theorem 5.1.

Theorem 5.1. *Under Assumption 9, 10, and 11, we have for* $t = 1, \dots, T, j = 1, \dots, N$

$$\hat{\boldsymbol{f}}_{\boldsymbol{t}} - \boldsymbol{H}_{NT}\boldsymbol{f}_{\boldsymbol{t}} = \frac{1}{NT}\left[\sum_{i=1}^{N}\sum_{s=1}^{T}\xi_{i,t}\boldsymbol{\Lambda}_i\boldsymbol{f}_s^{\top}\boldsymbol{H}_{NT}\boldsymbol{f}_s + \sum_{i=1}^{N}\sum_{s=1}^{T}\xi_{i,t}\xi_{i,s}\boldsymbol{H}_{NT}\boldsymbol{f}_s\right]\boldsymbol{D}^{-2}_{NT,r}$$

$$+ O_P\Bigg(\frac{\log(N)}{T} + \frac{k_\xi}{N} + \frac{\sqrt{\log(N)}}{\sqrt{NT}} + (NT)^{2/\zeta}\left(\frac{1}{\sqrt{NT}} + \frac{1}{T^{3/2}} + (NT)^{2/\zeta}\frac{1}{T^2}\right)\Bigg), \tag{5.7}$$

$$(\boldsymbol{H}_{NT}^\top)^{-1}\boldsymbol{\Lambda}_j - \hat{\boldsymbol{\Lambda}}_j = \frac{1}{T}\sum_{s=1}^{T} H_{NT}\boldsymbol{f}_s\xi_{j,s} + Error_j, \tag{5.8}$$

and

$$\begin{aligned} w_{j,t} &:= \hat{\xi}_{i,t} - \xi_{i,t} = \\ &= \boldsymbol{\Lambda}_j^\top \boldsymbol{H}_{NT}^{-1}\frac{1}{NT}\left[\sum_{i=1}^{N}\sum_{s=1}^{T}\xi_{i,t}\boldsymbol{\Lambda}_i\boldsymbol{f}_s^\top \boldsymbol{H}_{NT}\boldsymbol{f}_s + \sum_{i=1}^{N}\sum_{s=1}^{T}\xi_{i,t}\xi_{i,s}\boldsymbol{H}_{NT}\boldsymbol{f}_s\right]\boldsymbol{D}_{NT,r}^{-2} \\ &+ \boldsymbol{f}_t^\top \boldsymbol{H}_{NT}^\top \frac{1}{T}\left[\sum_{s=1}^{T} H_{NT}\boldsymbol{f}_s\xi_{j,s}\right] + Error_i, \end{aligned} \tag{5.9}$$

where

$$\begin{aligned} &\max_i |Error_i| = \\ &= O_P\Bigg(\frac{\log(N)}{T} + \frac{k_\xi}{N} + \frac{\sqrt{\log(N)}}{\sqrt{NT}} + (NT)^{2/\zeta}\left(\frac{1}{\sqrt{NT}} + \frac{1}{T^{3/2}} + (NT)^{2/\zeta}\frac{1}{T^2}\right)\Bigg). \end{aligned} \tag{5.10}$$

With the obtained representation for $\boldsymbol{w}_t$, we can analyze the estimation error of the second step. For this, note first that $\|1/T\sum_{t=1}^T(\boldsymbol{\xi}_t+w_t)(\boldsymbol{\xi}_t+w_t)^\top\|_{\max} \leq \|1/T\sum_{t=1}^T(\boldsymbol{\xi}_t)(\boldsymbol{\xi}_t)^\top\|_{\max}+2\|1/T\sum_{t=1}^T(w_t)(\boldsymbol{\xi}_t+w_t)^\top\|_{\max}+\|1/T\sum_{t=1}^T(w_t)(w_t)^\top\|_{\max}$ and the following Corollary 5.1.

Corollary 5.1. *Under the conditions of Theorem 5.1, we have the fol-*

lowing:

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_t\boldsymbol{\xi}_t^\top\right\|_{\max} =$$
$$= O_P\left(\frac{k_\xi}{N}+\frac{\log(N)}{T}+\frac{\sqrt{\log(N)}}{\sqrt{NT}}+(NT)^{2/\zeta-1}k_\xi+\frac{(NT)^{4/\zeta}}{T^2}\right),$$
$$\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_t\boldsymbol{w}_t^\top\right\|_{\max} =$$
$$= O_P\left(\frac{k_\xi}{N}+\frac{\log(N)}{T}+\frac{\sqrt{\log(N)}}{\sqrt{NT}}+(NT)^{2/\zeta}\left(\frac{1}{\sqrt{NT}}+\frac{1}{T^{3/2}}+(NT)^{2/\zeta}\frac{1}{T^2}\right)\right)$$

$\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_t\boldsymbol{\xi}_{t-1}^\top\|_{\max} = O_P(\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_t\boldsymbol{\xi}_t^\top\|_{\max})$, $\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_t\boldsymbol{w}_{t-1}^\top\|_{\max} = O_P(\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_t\boldsymbol{w}_t^\top\|_{\max})$.

As mentioned previously, if we just plug-in the rate for $\boldsymbol{w}_t$ we would obtain the slower rate of $O_P(\max(1/\sqrt{T},1/\sqrt{N}))$. With the results above we can establish bounds for the estimation error of the second step, see the following Theorem 5.2.

Theorem 5.2. *Under Assumption 9, 10, and 11 we have for* $j = 1,\ldots,N$

$$\begin{aligned}
&\|\hat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}^{(j)}\|_1 = \|\hat{\mathfrak{A}}-\mathfrak{A}\|_\infty = \\
&= O_P\Bigg(k\Bigg[\sqrt{\log(Np)/T}+(NpT)^{2/\zeta}/T+k\Bigg(\frac{k_\xi}{N}+\frac{\sqrt{\log(Np)}}{\sqrt{NT}} \\
&+(NpT)^{2/\zeta}\Bigg(\frac{k_\xi}{NT}+\frac{1}{\sqrt{NT}}+\frac{1}{T^{3/2}}+(NpT)^{2/\zeta}\frac{1}{T^2}\Bigg)\Bigg)\Bigg]^{1-q}\Bigg)
\end{aligned} \tag{5.11}$$

$$\|\hat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}^{(j)}\|_2 = O_P\Bigg(\sqrt{k}\Bigg[\sqrt{\log(Np)/T}+(NpT)^{2/\zeta}/T$$

$$
\begin{aligned}
&+ k\Bigg(\frac{k_\xi}{N} + \frac{\log(Np)}{T} + \frac{\sqrt{\log(Np)}}{\sqrt{NT}} \\
&+ (NpT)^{2/\zeta}\Big(\frac{k_\xi}{NT} + \frac{1}{\sqrt{NT}} + \frac{1}{T^{3/2}} + (NpT)^{2/\zeta}\frac{1}{T^2}\Big)\Big)\Bigg]^{1-q/2} \\
&+ k^{3/2}\Bigg[\sqrt{\log(Np)/T} + (NpT)^{2/\zeta}/T \\
&+ k\Bigg(\frac{k_\xi}{N} + \frac{\log(Np)}{T} + \frac{\sqrt{\log(Np)}}{\sqrt{NT}} \\
&+ (NpT)^{2/\zeta}\Bigg(\frac{k_\xi}{NT} + \frac{1}{\sqrt{NT}} + \frac{1}{T^{3/2}} + (NpT)^{2/\zeta}\frac{1}{T^2}\Bigg)\Bigg)\Bigg]^{(3-q)/2}\Bigg).
\end{aligned}
\tag{5.12}
$$

Let us have a closer look on the bound $\|\hat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^{(j)}\|_1$ and consider $N = T^a, p = T^b, k = T^c, k_\xi = T^d$ for some $a, b, c, d > 0$. Furthermore, let $\zeta \geq 4(1 + a + b)$. Then, we can simplify the error bound in O-notation and obtain $\|\hat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^{(j)}\|_1 = O_P(\log(T)^{(1-q)/2}T^{c-(1-q)/2} + T^{c+(1-q)(c+d-a)})$. That means a consistent estimation is obtained if $c < 1/2(1-q)$ and $c/(1-q) + c + d < a$. The first condition, i.e., this upper bound on the sparsity in relation to sample size T, is standard for approximately sparse models, see among others Corollary 2.4 in Geer (2016). In contrast, the second condition is not standard for approximately sparse models and appears due to the estimation error of the first step. This condition reflects the error occurring in factor models which are due to the introduced dependency of the VAR model not exact but only approximate. Hence, the time and cross-sectional dependency of the idiosyncratic component is quantified by k and k_ξ, i.e., c and d, and this dependency cannot be too strong in relation to the dimension such that it can be averaged out and a decent estimation of the common and idiosyncratic component can be obtained.

Note that if a sparsity is considered such that $\|\hat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^{(j)}\|_1 = o_P(1)$, then the terms with $k^{3/2}$ upfront in the error bound $\|\hat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^{(j)}\|_2$ are of higher order and negligible.

As mentioned, a major application of the dynamic factor model (5.1) is forecasting. For this, let $\boldsymbol{f}_{T+1}^{(1,p_f)} = \sum_{j=1}^{p_f} \boldsymbol{\Pi}_j^{(p_f)} \boldsymbol{f}_{T+1-j}$ be the linear one-step-ahead prediction based on $\boldsymbol{f}_T, \dots, \boldsymbol{f}_{T-p_f}$, where $\sum_{j=1}^{p_f} \boldsymbol{\Pi}_j^{(p_f)} \boldsymbol{\Gamma}_f(i-j) = \boldsymbol{\Gamma}_f(i), i = 1, \dots, p_f$ and $\boldsymbol{\Gamma}_f(i-j) = \mathbb{E}\boldsymbol{f}_{t+i}\boldsymbol{f}_{t-j}^\top$. Furthermore, since $\{\boldsymbol{\xi}_t\}$ follows a VAR(p) model, $\boldsymbol{\xi}_{T+1}^{(1)} = \boldsymbol{A}(\boldsymbol{\xi}_T^\top, \dots, \boldsymbol{\xi}_{T-p}^\top)^\top$ is the one-step-ahead prediction for the idiosyncratic component. That means, $\boldsymbol{X}_{T+1}^{(1,p_f)} = \boldsymbol{\Lambda}\boldsymbol{f}_{T+1}^{(1,p_f)} + \boldsymbol{\xi}_{T+1}^{(1)}$ is the joint one-step-ahead prediction for $\boldsymbol{X}_{T+1}$ with the prediction error $\text{Var}(\boldsymbol{X}_{T+1} - \boldsymbol{X}_{T+1}^{(1,p_f)}) = \boldsymbol{\Lambda}\text{Var}(\boldsymbol{f}_{T+1}^{(1,p_f)} - \boldsymbol{f}_{T+1})\boldsymbol{\Lambda}^\top + \Sigma_v$ and for a single variable j we have $\text{Var}(e_j^\top(\boldsymbol{X}_{T+1} - \boldsymbol{X}_{T+1}^{(1,p_f)})) = \boldsymbol{\Lambda}_j^\top \text{Var}(\boldsymbol{f}_{T+1}^{(1,p_f)} - \boldsymbol{f}_{T+1})\boldsymbol{\Lambda}_j + \boldsymbol{e}_j^\top \boldsymbol{\Sigma}_v \boldsymbol{e}_j$. If $\{\boldsymbol{f}_t\}$ follows a VAR(p_f) model, this simplifies to $\text{Var}(\boldsymbol{X}_{T+1} - \boldsymbol{X}_{T+1}^{(1,p_f)}) = \boldsymbol{\Lambda}\Sigma_u\boldsymbol{\Lambda}^\top + \Sigma_v$.

Since the parameters are unknown and the factors and idiosyncratic component are latent, this approach is unfeasible but the results of Theorem 5.1 and 5.2 help to obtain a feasible approach. For this we construct feasible counterparts of the prediction approach above, let $\hat{\boldsymbol{f}}_{T+1}^{(1,p_f)} = \sum_{j=1}^{p_f} \hat{\boldsymbol{\Pi}}_j^{(p_f)} \hat{\boldsymbol{f}}_{T+1-j}$ be the linear one-step-ahead prediction based on $\hat{\boldsymbol{f}}_T, \dots, \hat{\boldsymbol{f}}_{T-p_f}$, where $\sum_{j=1}^{p_f} \hat{\boldsymbol{\Pi}}_j^{(p_f)} \hat{\boldsymbol{\Gamma}}_f(i-j) = \hat{\boldsymbol{\Gamma}}_f(i), i = 1, \dots, p_f$ and $\hat{\boldsymbol{\Gamma}}_f(i-j) = 1/n\sum_{t=1+j}^{T-i} \hat{\boldsymbol{f}}_{t+i}\hat{\boldsymbol{f}}_{t-j}^\top$.
Furthermore, let $\hat{\boldsymbol{\xi}}_{T+1}^{(1)} = \hat{\boldsymbol{A}}(\hat{\boldsymbol{\xi}}_T^\top, \dots, \hat{\boldsymbol{\xi}}_{T-p}^\top)^\top$ be the one-step-ahead prediction for the idiosyncratic component. Then, $\hat{\boldsymbol{X}}_{T+1}^{(1,p_f)} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{f}}_{T+1}^{(1,p_f)} + \hat{\boldsymbol{\xi}}_{T+1}^{(1)}$ is the joint and feasible one-step-ahead prediction for $\boldsymbol{X}_{T+1}$. Since even though a high-dimensional time series system is considered the interest is often in the prediction of some key times series, we quantify in the following Theorem 5.3 the estimation error between the feasible and unfeasible approach for a single time series.

Theorem 5.3. *Under Assumption 9, 10, and 11 we have for* $j = 1, \ldots, N$

$$e_j^\top(\hat{\boldsymbol{X}}_{T+1}^{1,p_f} - \boldsymbol{X}_{T+1}^{(1,p_f)}) = O_P\Bigg(1/\sqrt{N} + k\Bigg[\sqrt{\log(Np)/T} + (NpT)^{2/\zeta}/T$$
$$+ k\Bigg(\frac{k_\xi}{N} + \frac{\sqrt{\log(Np)}}{\sqrt{NT}} + (NpT)^{2/\zeta}\Bigg(\frac{k_\xi}{NT} + \frac{1}{\sqrt{NT}} + \frac{1}{T^{3/2}}$$
$$+ (NpT)^{2/\zeta}\frac{1}{T^2}\Bigg)\Bigg)\Bigg]^{1-q}\Bigg).$$

In relation to the error bound for $\|\hat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^{(j)}\|_1$ derived in Theorem 5.2 only an additional $1/\sqrt{N}$ appears.

5.3.1 Estimation with Strong Idiosyncratic Components

In the error bounds in Theorem 5.2 and 5.3, the factor k_ξ/N plays an important role. $k_\xi = \|\mathrm{Var}(\boldsymbol{\xi}_t)\|_\infty$ quantifies the serial dependence of the idiosyncratic component. If this is large, the estimation in all steps suffers. Motivated by Generalized Least Squares (GLS) Boivin and Ng (2006) proposes to weight the data such that the serial dependence of the idiosyncratic component can be decreased. This approach is also denoted *generalized principal component analysis* and it is analyzed in more detail in Choi (2012). Let $\boldsymbol{W} \in \mathbb{R}^{N\times N}$ be a matrix of weights, then the factors are estimated using the weighted data $\boldsymbol{XW}$. Note that we have $\mathrm{Var}(\boldsymbol{XW}) = \boldsymbol{W\Lambda\Sigma}_F\boldsymbol{\Lambda}^\top\boldsymbol{W} + \boldsymbol{W\Gamma}_\xi(0)\boldsymbol{W}^\top$. Hence, the factors can be estimated by a PCA of $\boldsymbol{XW}$ whereas the loadings are obtained by regressing $\boldsymbol{X}$ onto the estimated factors. Since non-diagonal weighting schemes are seldom feasible without sparsity constraints, Boivin and Ng (2006) suggest different diagonal weighting schemes. With the additional assumption that the $\boldsymbol{\Sigma}_v$ – the variance matrix of the idiosyncratic innovation $\boldsymbol{v}_t$ – is sparse, we suggest to use the VAR structure of the idiosyncratic component to obtain a more refined weighting scheme.

To elaborate, we have that $\operatorname{Var}(\boldsymbol{\xi}_t) = \boldsymbol{\Gamma}_\xi(0) = \sum_{j=0}^{\infty} \boldsymbol{B}^{(j)} \boldsymbol{\Sigma}_v (\boldsymbol{B}^{(j)})^\top$, where $(\boldsymbol{B}^{(j)})_{r,c} = (\mathfrak{A}^j)_{r,c}, r, c = 1, \ldots, N$. Hence, $\boldsymbol{\Gamma}_\xi(0)$ is given by $\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(p)}, \boldsymbol{\Sigma}_v$ and it can be estimated by plugging in estimators, see among others Theorem 5 in Krampe and Paparoditis (2021). Let us denote this estimator as $\hat{\boldsymbol{\Gamma}}_\xi(0)$. Depending on whether sparsity constraints on $\boldsymbol{\Sigma}_v$ or $\boldsymbol{\Sigma}_v^{-1}$ are more realistic, estimators are given by thresholding of the empirical covariance matrix (Bickel and Levina, 2008; Cai and Liu, 2011) or by component-wise regularized regression (Friedman et al., 2008; Cai, Liu, and Zhou, 2016; Cai, Ren, et al., 2016). The weighting matrix is then given as $\boldsymbol{W} = \hat{\boldsymbol{\Gamma}}_\xi(0)^{-1/2}$. Consequently, the "new" k_ξ is given by $\|\hat{\boldsymbol{\Gamma}}_\xi(0)^{-1/2} \boldsymbol{\Gamma}_\xi(0) \hat{\boldsymbol{\Gamma}}_\xi(0)^{-1/2}\|_\infty$ which can be considerably smaller if the used estimators give reasonable results. Since the weighting leads also to a new estimation of the idiosyncratic component, it might be helpful to apply this approach more than once.

5.3.2 Similarity and Differences to Low-Rank plus Sparse Models

As mentioned the low-rank plus sparse VAR model discussed in Basu, Li, et al. (2019) as well as the linear dynamical system discussed in Lin and Michailidis (2019) are related to the model proposed here. We now stress out the similarities and differences of these models beginning with the model of Basu, Li, et al. (2019). The low-rank plus sparse VAR model of order p is given by $\boldsymbol{x}_t = \sum_{j=1}^{p} \boldsymbol{\Theta}^{(j)} \boldsymbol{x}_{t-j} + \boldsymbol{\varepsilon}_t,\ \boldsymbol{\Theta}^{(j)} = \boldsymbol{L}^{(j)} + \boldsymbol{S}^{(j)}, \operatorname{rank}(\boldsymbol{L}^{(j)}) = r \ll N$, where $\boldsymbol{\varepsilon}_t$ is some white-noise process, L is a low-rank matrix, and $\boldsymbol{S}^{(j)}$ possesses some sparsity structure. The low-rank matrix takes here the role of the common component, see also Bai and Ng (2019). Thus, this approach also combines a dense and a sparse approach. However, there are two major differences to the approach presented in this Chapter. First, note that while the low-rank plus sparse VAR model is some special form of a VAR(p) model, a dynamic factor model is instead in general a VAR(∞) process even if the factors and idiosyncratic components follow finite order VAR processes. Second and most important, with the approach presented here we can derive estimation error bounds for a singe time series, see Theorem 5.3.

This is in contrast to the results derived in Basu, Li, et al. (2019). They impose sparsity constraints on vec$(\boldsymbol{S}^{(j)})$[2] and they do not estimate the VAR system row-wise as in (5.4). Instead, all regression equations are combined using the Frobenius norm, the VAR slope matrices are considered as a sum of two matrices where the first matrix is regularized using the nuclear norm – this imposes a low-rank structure – and the second matrix is regularized using the ℓ_1 norm on the vectorized matrix – this imposes a sparse structure. They derive error bounds only regarding the Frobenius norm. That means they consider only the overall estimation error. In connection with the sparsity constraints on vec$(\boldsymbol{S}^{(j)})$ this is too restrictive or too less detailed for the row-wise estimation error which is helpful for a single forecast. For a more detailed discussion of the different sparsity concepts and their implication regarding estimation error bounds we refer to Section 2 in Krampe and Paparoditis (2021).

For the model discussed in Lin and Michailidis (2019) note first that a dynamic factor model $\boldsymbol{x}_t = \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{\xi}_t$ whose idiosyncratic component follows a VAR(p) model $\boldsymbol{\xi}_t = \sum_{j=1}^{p} \boldsymbol{A}^{(j)} \boldsymbol{\xi}_{t-j} + \boldsymbol{v}_t$ can be written as $\boldsymbol{x}_t = \sum_{j=1}^{p} \boldsymbol{A}^{(j)} \boldsymbol{\Lambda} \boldsymbol{f}_{t-j} + \sum_{j=1}^{p} \boldsymbol{A}^{(j)} \boldsymbol{x}_{t-j} + \boldsymbol{v}_t$. The component $\sum_{j=1}^{p} \boldsymbol{A}^{(j)} \boldsymbol{\Lambda} \boldsymbol{f}_{t-j}$ can be considered as the common component of general dynamic factor model as in Forni, Hallin, Lippi, and Reichlin (2000) and it is low-rank. Lin and Michailidis (2019) consider that the matrices $\boldsymbol{A}^{(j)}, j = 1, \ldots, p$ are sparse and they impose sparsity constraints on vec$(\boldsymbol{S}^{(j)})$. Furthermore, they combine all regression equations using the Frobenius norm and the low-rank part is handled by regularization of its nuclear norm. Similarity to Basu, Li, et al. (2019), they derive error bounds only regarding the Frobenius norm. That means for a forecast of a single time series the same drawbacks described above apply. Let us note here that it is not clear if the approach of handling the low-rank part by regularization of its nuclear norm can be also done equation by equation such that error bounds more helpful for a forecast of a single time series can be derived.

[2] Basu, Li, et al. (2019) consider also a group-sparse structure for $S^{(j)}$. For this sparsity concept the discussion is quite similar.

5.4 Number of Factors, Lag-length and Penalty Tuning

5.4.1 Number of Factors

The seminal work of Bai and Ng (2002) introduced a focus over the correct specification of the number of factors, which were up to that point mostly assumed, rather than data-driven. The idea for how to tackle estimation of the number of common factors is induced by the well known fact that a certain amount of eigenvalues of the covariance matrix of the data diverges to infinity (i.e.,the common factors), whereas the remaining ones stay bounded. As a consequence, if one finds a threshold able to clearly separate among those finite and infinite eigenvalues, the problem is solved. Bai and Ng (2002) designed precisely this, namely an information criteria able to threshold the diverging eigenvalues. Their key contribution lies in the fact that they allow both the cross-sectional dimension as well as the time dimension to diverge. The immediate consequence is the exclusion from the pool of candidate techniques aimed at consistently estimate the number of factors, of any standard information criteria like AIC or BIC, only depending on either one of the dimensions. In fact, their framework calls for the penalty for overfitting to be a function of both N and T. Their approach in estimating the common factors is non-parametric through the method of asymptotic principal components and their asymptotic results yields consistent estimation of the number of factors. However, in practice, Bai and Ng (2002) criterion suffers from a penalty identification issue which can return non-robust results as the number of factors can be overestimated or underestimated. As observed in Hallin and Liška (2007) and Alessi et al. (2010) (HLA henceforth), for respectively the cases of dynamic and static factor models, the consistency of the estimated number of factors via the Bai and Ng (2002) criterion still holds if the penalty parameter is multiplied by an arbitrary real, positive constant c. While this is asymptotically elegant, in finite samples (both N and T) can result in arbitrary large or arbitrary small values of the penalty, thus gravely affecting the results. To bypass this issue, HLA proposed to pre-multiply the penalty function by a positive real number c. In other words, by letting the criterion to be dependent on c and by splitting the sample size in sub-batches then the number of factors calculated over the batches becomes a monotonic function of c. The tuning of c then allows for either no penalization when $c = 0$, underpenalization when c is positive but small and overpenalization when c is positive but large. It follows that c is optimized whenever the value of c let the number of factors to be a stable function over the batch.

Alternatively to thresholding the eigenvalues, one could compute the ratio of adjacent eigenvalues as this is also bound to diverge. Among others, Onatski (2009) followed this route, showing the asymptotic distribution of their proposed test statistics to be a function of the Tracy-Widom distribution.

While the above mentioned criteria to choose the number of factors are standard and often computationally appealing, they do rely on a set of assumptions over the factors, their loadings and the idiosyncratic errors. An alternative methodology to estimate the number of factors is proposed in Trapani (2018) for the case of a static approximate factor model. There, a sequential, randomized test for the jth eigenvalue being divergent or bounded is proposed. The estimation of the factors is then found to be robust to a wide variety of data generating processes, including those affected by serial and cross-sectional dependence and also to the presence of weak factors.

5.4.2 Lag-length

Until now we assumed the lag-length p to be given. Clearly, in practice, p needs to be estimated. Theoretically, one can safely assume the VAR lag-length p to be reasonably small in a high-dimensional framework (see e.g., Hecq et al., 2016). The reason is to be found in deriving the VAR(p) final equation representation (Zellner and Palm, 1974). Consider the VAR equation (5.3) for the idiosyncratic components in terms of the lag-operator L, such that for $\boldsymbol{A}(L) = (\boldsymbol{I} - A^{(1)}L - \ldots - A^{(p)}L^p)$:

$$\boldsymbol{A}(L)\boldsymbol{\xi}_t = \boldsymbol{v}_t. \tag{5.13}$$

Pre-multiplying both sides of (5.13) by $\boldsymbol{A}^{\text{adj}}(L) = \det(\boldsymbol{A}(L))\boldsymbol{A}(L)^{-1}$ i.e.,the adjoint of the matrix polynomial, one obtains

$$\det(\boldsymbol{A}(L))\boldsymbol{\xi}_t = \boldsymbol{A}^{\text{adj}}(L)\boldsymbol{v}_t. \tag{5.14}$$

Each cross-sectional equation in (5.14) then follows an ARMA$(Np, (N-1)p)$ which is of maximal orders already e.g.,for a setting as $N = 100$ and $p = 2$.

Practical ways to estimate a (global) lag-length are introduced in Chapter 3 where we marginalize the VAR system into a sequence of univariate AR(p) processes, and select the lag-length by minimizing an approximated Bayesian Information criterion (BIC) on the residual covariance matrix. Such procedure is effectively an empirical upper-bound since the VAR system is diagonalized to avoid the dimensionality issue. As a consequence, the lag-length is actually estimated only on the autoregressive component of each equation, which will naturally tend to upper-bound the true lag. However, in practice this approach selects the right lag-length most of the time with only seldom minor over-estimations in large systems. Consistency of the BIC up to a slowly diverging constant C_N (as long as $C_N N\log(T)/T \to 0$) in the penalty term has been proved in Wang, Li, and Leng (2009). There, they show under a set of technical assumptions on: the divergence speed of the model dimension ($\limsup(N/T^{\alpha}) < 1$ for $\alpha < 1$), the size of the non-zero coefficients ($\sqrt{[T/C_T N\log(T)]}\liminf_{T\to\infty}(\min_{j\in S}|\boldsymbol{\beta}_j|) \to \infty$) and minimum eigenvalue of the covariance matrix to be bounded away from zero, that this only slightly modified BIC can identify the true model consistently even when the dimensions diverge.

Alternatively, one can embrace the high-dimensionality and directly apply regularizations of the lasso type. Note how this only shifts the model selection problem from a BIC applied over estimated AR(p) residuals as in Chapter 3, to the choice of the tuning parameter in the regularization technique. The rationale of the latter approach is clear: as the lasso should shrink and eventually set to zero the coefficients of those irrelevant parameters, then one should have that the elements of the lasso-estimated $\hat{A}^{(j)} = 0$ for all $j > p$ where p is the true lag-length. Selection consistency of these shrinkage estimators crucially relies on how one tunes their penalty parameter. Although widely employed in practice, the generalized cross validation method does not consistently recover the true model even in fixed dimension settings (see e.g.,Wang, Li, and Tsai (2007) and Wang and Leng (2007) for examples with SCAD and Adaptive lasso). Wang, Li, and Leng (2009) shows again that their slightly modified BIC, when employed to tune the penalty parameter in SCAD and lasso, renders these shrinkage estimators selection-consistent.

In light of the decomposition shown in (5.14), there seems to be not much gain, both computationally and in terms of possible erratic behaviors due to high-correlations, to use these shrinkage estimators for the purpose of lag-length selection. This is true especially if one is more concerned with inference than forecasting and, consequently, is more lenient over the usual assumption of a unique lag-length for the whole VAR system (cf. short range dependence). One interesting exception is the hierarchical penalties of Nicholson, Wilms, et al. (2020), these include the notion of lag selection into a convex regularizer and they can be used on a set of values for p, possibly varying the lag-length over different variables. A group lasso with nested groups guarantees that the sparsity pattern of lag coefficients correctly mimics the VAR structure.

5.4.3 A Combined Approach for Single Time Series

In light of the previous paragraphs, we seek for a unified procedure able at the same time to consistently estimate the lag-length as well as the number of factors. For a given lag-length and number of factors, the penalty parameter can be chosen with the approaches discussed in the next paragraph and we take this as granted here. Furthermore, we would like to put the focus more on the objective of forecasting single time series of the system. In doing so we would also like to allow that the lag-length or number of factors may differ across the time series.

The reason for this is that large data sets come as a – in some sense arbitrary – collection of series and it is most likely that some series are not driven by factors or a small lag-length is sufficient. Since parameters can be zero, this is of course all contained in a model which number of factors and lag-length is given by maximum over the individual series. However, this can make a difference in finite samples.

For this, we consider that the factors are driven by a VAR model that is $\boldsymbol{f}_t = \sum_{j=1}^{p_f} \boldsymbol{\Pi}_j \boldsymbol{f}_{t-j} + \boldsymbol{v}_{t-j}$. That means we have two lag-lengths to chose: p and p_f. The one-step ahead forecast error of model (5.1) for the j component is given by

$$\operatorname{Var}\Big(x_{i,t} - \sum_{j=1}^{p_f} \boldsymbol{\Lambda}_i^\top \boldsymbol{\Pi}_j \boldsymbol{f}_{t-j} - \sum_{j=1}^{p} e_i^\top \boldsymbol{A}_j \boldsymbol{\xi}_{t-j}\Big).$$

If we do not treat the estimation of the factors and idiosyncratic components as additional parameters, we have the parameters $\boldsymbol{\Lambda}_i\boldsymbol{\Pi}_1, \dots, \boldsymbol{\Lambda}_i\boldsymbol{\Pi}_{p_f} \in \mathbb{R}^r, e_i^\top \boldsymbol{A}_1, \dots, e_i^\top \boldsymbol{A}_p \in \mathbb{R}^N$. Note that $e_i^\top \boldsymbol{A}_1, \dots, e_i^\top \boldsymbol{A}_p$ are sparse and we treat $\boldsymbol{\Lambda}_i\boldsymbol{\Pi}_j$ as r-dimensional vectors. That means in total we have $rp_f + \sum_{j=1}^{p} \|e_i^\top \boldsymbol{A}_j\|_0$ parameters for the jth component. Following the approach of Wang, Li, and Leng (2009) with a modified BIC and C_T denotes a diverging series, this motivates the following information criteria

$$\begin{aligned} IC_{T,N} := \min_{r,p,p_f} \log \frac{1}{T} \sum_{t=1+\max(p,p_f)}^{T} &\left(x_{i,t} - \sum_{j=1}^{p_f} \hat{\boldsymbol{\Lambda}}_i^\top \hat{\boldsymbol{\Pi}}_j \hat{\boldsymbol{f}}_{t-j} - \sum_{j=1}^{p} e_i^\top \hat{\boldsymbol{A}}_i \hat{\boldsymbol{\xi}}_t^{(r)} \right)^2 \\ &+ (rp_f + \sum_{j=1}^{p} \|e_i^\top \hat{\boldsymbol{A}}_j\|_0) \frac{\log(T)}{T} C_T. \end{aligned} \tag{5.15}$$

In practice the minimum is evaluated over a finite grid. That means one sets maximal number of factors $r_{\max}$ and maximal lag-lengths $p_{\max}, p_{2,\max}$. If one sets $r_{\max} = 0$ or $p_{\max} = 0$, this criteria can also be used to fit plain sparse VAR models or plain factor models, respectively. The series C_T can be diverging very slowly and Wang, Li, and Leng (2009) suggest for instance, $\log(\log(T))$. We would like to consider the diverging dimension and follow a similar route as Bai and Ng (2002), Hallin and Liška (2007), and Alessi et al. (2010). So we set $C_T = c\frac{\log(NT/(N+T))}{\log(T)}$ and $c = 1/2$.

5.4.4 Penalty Tuning

Another important aspect to consider within the framework of ℓ_1-regularizations is the choice of the tuning parameter λ. The latter should be set in order to balance between the fit of the model and its complexity, thus trading off bias with variance. Whenever the tuning parameter is large in its magnitude, the consequence is strong variable selection, i.e., many potentially relevant variables might be set to zero by the regularization technique (e.g lasso), thus implying a larger estimation bias. In parallel, when instead $\lambda \approx 0$, no variable selection is performed and thus regularization techniques such as lasso converge in the limit to the standard OLS estimator. Among the most popular techniques to tune λ is cross-validation (CV). While CV have seen a surge of applications in statistics in the last ten years, it can suffer of some shortcomings. First, it is often computationally demanding, especially in high dimensions, given it has to recursively train and validate on batches of the sample. Second, it needs to be adapted for different data. For instance, when working with time series, CV needs to be adapted to avoid temporal dependence disruptions when defining the folds (see e.g. Bergmeir et al., 2018). In fact, a rolling K-fold cross validation needs to be used in order to gradually train the series avoiding to lose their dependence. Chetverikov et al. (2020) showed that K-fold CV applied to the lasso has nearly optimal rates of convergence in ℓ_1, ℓ_2 prediction norms. However, the technique tends to render small values of λ which in turn imply less variable selection, thus making it not a particularly favorable method, especially in very large dimensions where more shrinkage would be required.

Alternatively, a fast and reliable way of tuning λ is by minimizing an information criterion (IC). Let $\boldsymbol{\xi}^v_{t,S}$ be the subvector containing those columns of $\boldsymbol{\xi}_t^v$ belonging to the set S. Let further $\hat{S}$ be the active set identified by the lasso for a given λ. Then the value λ^{IC} chosen by information criteria is found as

$$\lambda^{IC} = \arg\min_{\lambda} \left(\ln \left(\frac{1}{T-p+1} \sum_{t=p+1}^{T} \left(\xi_{j,t} - \sum_{j=1}^{p} \boldsymbol{\beta}_{S(\lambda)}^{\top} \boldsymbol{\xi}_{t-j,S(\lambda)}^{v} \right)^2 \right) \right.$$
$$\left. + \left(\frac{1}{T-p+1} \right) C_T df \right),$$

where df represents the degrees of freedom after the penalization, i.e., the cardinality of the estimated active set. C_T is the penalty specific to each criterion, where the most popular choices are: $C_T = 2$, the Akaike information criterion (AIC) by Akaike (1974); $C_T = \log(T)$, the Bayesian information criterion (BIC) by Schwarz (1978). As for the case of the lag-length selection in Remark 5.4.2, the slight modification of the BIC proposed in Wang, Li, and Leng (2009) also holds for penalized estimators as the lasso, thus making it consistent asymptotically in both N and T.

Yet another approach in tuning λ is a theoretical one (see e.g. Bickel, Ritov, et al., 2009; Belloni and Chernozhukov, 2013; Belloni, Chernozhukov, and Wang, 2011). Namely, the tuning parameter has to be set to upper bound the gradient of the criterion function (i.e., the score), thus introducing bias towards zero to reduce the variance. Then, it is enough to require with high probability that $\lambda \geq c||\boldsymbol{\xi}_t'\boldsymbol{v}||_\infty/T$, where c is an absolute constant and $c||\boldsymbol{\xi}_t'\boldsymbol{v}||_\infty/T$ is often referred to as the *effective noise*. In fact, any high-probability bound on the effective noise can be used as lower bound for λ. Since $\boldsymbol{v}$ is unknown, several Gaussian approximations has been proposed in the literature (see inter alias: Chernozhukov, Chetverikov, et al., 2013; Chernozhukov, Chetverikov, et al., 2014; Zhang and Wu, 2017). With normality of the innovations the choice of the tuning parameter reduces to simple expressions for λ depending on the Gaussian CDF. Similarly, Belloni and Chernozhukov (2013), Belloni, Chernozhukov, and Hansen (2011a), and Chernozhukov, Hansen, et al. (2016) use penalty loadings to first-order self-normalize the the lasso problem and hence applying moderate deviation theory results (see Jing et al., 2003) to bound deviations of the maximal element of the score vector. The only disadvantage of such theoretically sounding alternatives is that the practitioner can only strengthening or weakening the magnitude of λ through arbitrary increases or decreases of step $\tilde{\epsilon}$ of the universal constant(s) c present in the derived expressions for λ, i.e., $c \pm \tilde{\epsilon}$. In other words, they depend on some free parameters and hence are not entirely data-driven. An overview of the performances of these different choices of λ can be found in Chapter 2 where we compare these under the setting of Granger causality testing in high-dimensional stationary VAR models. A notable recent exception is represented by Lederer and Vogt (2020). They design a bootstrap-based estimator of the quantiles of the effective noise which attains optimal finite-sample guarantees and it does not depend on any free parameters.

5.5 Numerical Results

5.5.1 Simulation Set-up & Data Generating Processes

All results presented in this section are based on implementations in *R* (R Core Team, 2020). In the simulation setup, we generate the data generating processes (DGPs) at random and consider the following model class: $\boldsymbol{x}_t = \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{\xi}_t, \boldsymbol{f}_t = \sum_{j=1}^{p_f} \boldsymbol{\Pi}^{(j)} \boldsymbol{f}_{t-j} + \boldsymbol{u}_t, \boldsymbol{\xi}_t = \sum_{j=1}^{p} \boldsymbol{A}^{(j)} \boldsymbol{\xi}_{t-j} + \boldsymbol{v}_t$. The innovations $\{\boldsymbol{u}_t\}, \{\boldsymbol{v}_t\}$ are generated as Gaussian processes and $\boldsymbol{\Sigma}_u = \mathrm{Var}(\boldsymbol{u}_t)$ is generated as a positive definite matrix with eigenvalues in the range 1 to 10 using the implementation of the package *clusterGeneration* (Qiu and Joe., 2020). If not denoted otherwise, sparsity of a matrix is obtained by setting entries – beginning with the absolute smallest values – to zero such that the specified amount of sparsity is obtained. Furthermore, we consider the following specifications:

- The number of factors is given by $r \in \{0, 2, 4, 6\}$.
- The sample is given by $T \in \{100, 200\}$.

- The dimension is given by $N \in \{50, 100, 250\}$.
- The lag-length of the VAR driving the factors is given by $p_f \in \{0, 1, 2\}$. The slope matrices are generated at random and the maximal absolute eigenvalue of the stacked VAR matrix is 0.8.
- The lag-length of the VAR driving the idiosyncratic component is given by $p \in \{0, 1, 3\}$. The slope matrices are generated at random with a row-wise and column-wise sparsity of $k \in \{5, 10, min(N, 100)\}$ and the maximal absolute eigenvalue of the stacked VAR matrix is 0.8.
- $\boldsymbol{\Sigma}_v = \text{Var}(\boldsymbol{v}_t)$ is generated as a positive definite matrix with eigenvalues in the range 1 to 10 and sparsity of $k_\Sigma \in \{N/10, N\}$.
- The loadings $\Lambda \in \mathbb{R}^{N \times r}$ are generated by random sampling from a Uniform$[-1, 1]$ distribution with a column-wise sparsity of $k_\Lambda \in \{N, N/2, N/2^*\}$. $N/2^*$ refers to a setting in which the lower left and upper right part are zero. For this setting, also the the lower left and upper right part of $\boldsymbol{\Pi}^{(j)}, j = 1, \ldots, p_f$ and $\boldsymbol{\Sigma}_u$ is set to zero.

Note that a sparsity of N implies no sparsity. Dropping unnecessary combination, e.g., varying the sparsity for $p = 0$, we end up in 2352 different set ups for the DGP. We run each set up 100 times. To evaluate the performance, we consider the average one-step ahead prediction error of the first ten time series based on T observation of the original processes and evaluated at 1000 time points. That is $MSFE_{\boldsymbol{x}} = \frac{1}{10}\sum_{i=1}^{10}\left[\frac{1}{1000}\sum_{t=1}^{1000}(\hat{x}_{i,T+t}^{(1)} - x_{i,T+t}^{(1)})^2\right]$. We consider the following models to predict:

> *DFMsVAR*: The approach presented in this Chapter, i.e., a DFM with a VAR for the factors and a sparse VAR for the idiosyncratic component. The number of factors and lag-length are chosen by the information criteria of Section 5.4.3. The sparse VAR is estimated by a row-wise adaptive lasso and the penalty parameter is chosen by BIC.

sVAR: A sparse VAR which is estimated by a row-wise adaptive lasso and the penalty parameter is chosen by BIC. The lag-length is chosen by the information criteria of Section 5.4.3 and the maximal number of factors is set to zero.

DFMAR: A DFM with a VAR for the factors and univariate AR for the idiosyncratic component. The number of factors and lag-length for the VAR is chosen by the information criteria of Section 5.4.3. The lag-lengths of the univariate ARs are chosen by AIC.

DFM(ABC)AR: A DFM with a VAR for the factors and univariate AR for the idiosyncratic component. The number of factors is chosen by information criteria of Alessi et al. (2010), the lag-length for the VAR is chosen by BIC and The lag-lengths of the univariate ARs are chosen by AIC.

AR: Univariate ARs which lag-length is chosen by BIC.

Mean: A simple mean forecast.

In the following, we present the MSFE-results in relation to the MSFE of *DFMsVAR* meaning values larger than 1 indicate a performance worse than *DFMsVAR* and values smaller than 1 vice versa. The overall performance is summarized in Table 5.1. *DFMAR* and *DFM(ABC)AR* differ only in their factor selection criteria. Their performance not only overall but also in most set ups is quite similar which is why we do not present results for *DFMAR* in the following.

sVAR	*DFMAR*	*DFM(ABC)AR*	*AR*	*Mean*
1.05	1.25	1.26	1.34	1.39

Table 5.1: Overall performance measured in MSFE and in relation to *DFMsVAR*.

The relative performance over all 2352 different DGP set ups is displayed in Figure 5.1. Each dot represents the relative MSE for one DGP set up averaged over the runs. The set ups are sorted by lag-length of the idiosyncratic part p, lag-length of the factors p_f, and sparsity. The obtained groups are highlighted by vertical bars and the specific parameter values are given at the bottom of the figure. This sorting is chosen since these specification parameters matters the most in the sense that the results can differ substantially among different specification of the parameter values.

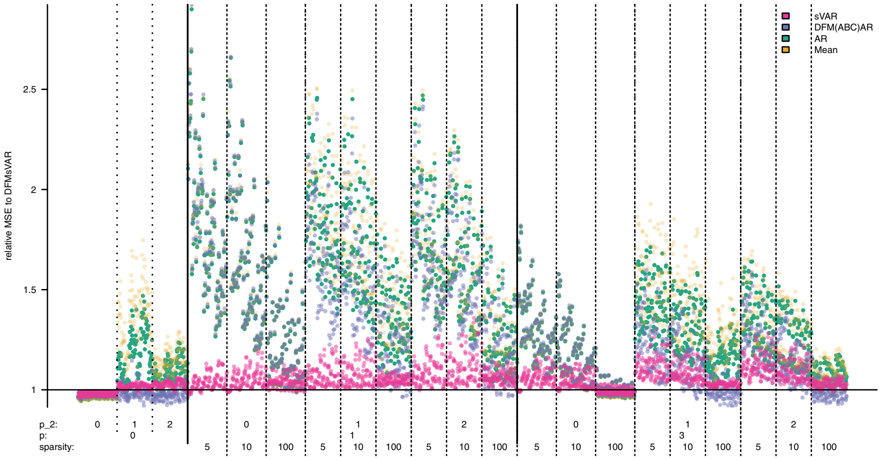


Figure 5.1: The relative performance over all 2352 different DGP set ups. Each dot represents the relative MSE for one DGP set up. The set ups are sorted by lag-length of the idiosyncratic part p, lag-length of the factors p_f, and sparsity. The obtained groups are highlighted by vertical bars and the specific parameter values are given at the bottom of the figure. Note that a sparsity of 100 implies in principle here no sparsity at all.

In the first big block, i.e., $p = 0$, the idiosyncratic component possesses no auto-correlation and *DFM(ABC)AR* performs non-surprisingly the best. However, the outperformance is quite moderate and even the plain lasso approach, *sVAR*, is only slightly worse. In contrast, if the idiosyncratic component possesses auto-correlation and is sparse (sparsity $\in \{5, 10\}$), *DFM(ABC)AR* is considerably worse than the *sVAR*.

Here, *DFMsVAR* outperforms all other approaches. This behavior does not change if the factors add additional auto-correlation or not. What matters is the sparsity of the idiosyncratic component. If it is not sparse anymore (sparsity = 100), all approaches do not perform well as indicated by the fact that the mean is only slightly outperformed. In the case when no factors and only a non-sparse component are present ($p_f = 0, p = 3$, sparsity= 100), all approaches perform as bad as the mean despite the non-sparse component bringing a strong auto-correlation. Furthermore, in the presence of factors and a non-sparse component, *DFM(ABC)AR* outperforms the lasso and in some cases also *DFMsVAR*.

5.6 Conclusion

We blend the dense dimensionality reduction of factor models with the one of sparsity-inducing high-dimensional VARs. Hence, we propose a *dynamic factor model* whose factors and relative loadings are estimated via standard principal components while its idiosyncratic components are assumed to follow a high-dimensional sparse VAR model and are thus estimated via ℓ_1-norm regularization techniques such as the lasso. The estimation is articulated in two steps: first the factors and their loadings are estimated via standard singular value decomposition and the estimated idiosyncratic components are obtained as estimated residuals. Second, an adaptive lasso is estimated on the previously obtained idiosyncratic components. As the second step is performed on the estimated idiosyncratic term from step 1, in order to derive the consistency of the lasso we first derive a simple representation of the idiosyncratic estimation error which helps to obtain sharper rates for the deviation bound in the second step. The simple representation is derived from the decomposition idea in Bai and Ng (2020) which also allows to obtain an expression for the error between estimated factors/loadings and a rotated version of the true factors/loadings and hence derive as well the order of their estimation errors. Through this blended approach we are able to disentangle the dependence among the factors, i.e., the diverging eigenvalues of the covariance matrix, and the dependence among the idiosyncratic components, i.e., the bounded ones, while allowing for both cross-sectional and time dependence in the idiosyncratic term. In order to choose on the one hand the number of factors and on the other the lag-length of the VAR, we also propose a unified procedure able simultaneously to estimate both. We consider the factors being driven by a VAR(p_f) while idiosyncratic components follow a VAR(p) and we then set up an information criteria minimizing the one-step ahead forecast error of the model over a grid of both p, p_f and the number of factors. The composite penalty of the criterion allows the joint selection of lags and number of factors. We corroborate our results with a thorough simulation exercise in which several data generating processes are considered both for different sparsity patterns as well as different specifications of the numbers of factors and lag-lengths. We compare the performances of our proposed method with several workhorse forecasting models in the literature and find that the procedure proposed here performs well.

Appendix A Proofs and additional Lemmas

In order to quantify the dependence of the stochastic processes, we use the concept of functional dependence, see Wu (2005), and concentration inequalities derived under this concept of dependence, see among others Liu et al. (2013) and Wu and Wu (2016). In the following remark 5.1 we summarize the main notation of this dependence concept.

Remark 5.1 (Functional Dependence Measure)**.** Let $Y_{t;i} = G_i(\varepsilon_t, \varepsilon_{t_1}, \dots,), i = 1, \dots, N, t \in \mathbb{Z}$, be some process generated causally by the i.i.d. processes $\{\varepsilon_t\}$ for some function $G = (G_1, \dots, G_N)$. Furthermore, denote by $Y'^{(k)}_{t;i} = G_i(\varepsilon_t, \varepsilon_{t-1}, \dots, \varepsilon_{t-k+1}, \varepsilon'_{t-k}, \varepsilon_{t-k-1}, \varepsilon_{t-k-2}, \dots)$ the process where ε_{t-k} is replaced by an i.i.d. copy ε'_{t-k}. We follow Wu (2005) and Wu and Wu (2016) and define the physical/functional dependence coefficients in the following way. Let $\|\xi_{i,t}\|_{E,q} := (\mathbb{E}\,|\xi_{i,t}|^q)^{1/q} < \infty, q \geq$

1. Furthermore, let the functional dependence measure be defined as $\delta_{k,q,i} = \|Y_{0;i} - Y_{0;i}^{\prime(k)}\|_{E,q}, k \geq 0$. In order to account for the dependence in the process $Y_{;i}$ let $\Delta_{m,q;i} = \sum_{k=m}^{\infty} \delta_{k,q;i}$ such that the dependence adjusted norm is defined as $\|Y_{;i}\|_{q,\alpha} = \sup_{m\geq 0}(m+1)^{\alpha}\Delta_{m,q;i}$. As we work in a high-dimensional setting, in order to take this into account we need a uniform dependence adjusted norm $\Psi_{q,\alpha} = \max_{1\leq i\leq N} \|Y_{\cdot;i}\|_{q,\alpha}$, and an overall dependence adjusted norm $\Upsilon_{q,\alpha} = (\sum_{i=1}^{N}(\sup_{m\geq 0}(m+1)^{\alpha}\Delta_{m,q;i})^{q})^{1/q}$. Furthermore, define for the N dimensional stationary process $Y_{t;i}$ the $\mathcal{L}^{\infty}$ functional dependence measure with its corresponding dependence adjusted norm: $\omega_{k,q} = \|\|Y_{t;i} - Y_{t;i}^{\prime(k)}\|_{\infty}\|_{E,q}$, $\|\|Y_{\cdot}\|_{\infty}\|_{q,\alpha} = \sup_{m\geq 0}(m+1)^{\alpha}\Omega_{k,q}$ for $\Omega_{m,q} = \sum_{k=m}^{\infty}\omega_{k,q}$. Finally, let $\nu_q = \sum_{j=1}^{\infty}(j^{q/2-1}\omega_{k,q})^{1/(q+1)}$.

Assumption 9 implies that $\|e_i B^{(j)}\|_2 \leq M\rho^j$. Hence, it follows by Example 3 in Wu and Wu (2016) and the moment condition in Assumption 10 that $\max_j \|\{\xi_{j,t}\}\|_{\zeta,\alpha} < \infty$ for all $\alpha > 0$. Since $\{\boldsymbol{f}_t\}$ is a linear processes of fixed dimension r, we also have $\max_j \|\{\chi_{j,t}\}\|_{\zeta,\alpha} < \infty$ for all $\alpha > 0$. Hence, we have by the Minkowski-inequality $\max_j \|\{x_{j,t}\}\|_{\zeta,\alpha} < \infty$, see also the Proof of Proposition 5 in Forni, Hallin, Lippi, and Zaffaroni (2017). Additionally we have by the Cauchy-Schwarz-inequality for some $q > 2$, $\max_{j,i} \|\{\xi_{j,t}f_{i,t}\}\|_{q,\alpha} \leq C(\max_j \|\{\xi_{j,t}\|_{2q,\alpha} + \max_j \|\{f_{i,t}\|_{2q,\alpha} + \max_j \|\{\xi_{j,t}\|_{2q,\alpha} \max_j \|\{f_{i,t}\|_{2q,\alpha})$, where C is some constant depending on q only.

Lemma 5.1. *Let C_1, C_2, C_3 be constants depending only on q and α. Under Assumption 9, 10, 11 we have the following:*

A) $\|\boldsymbol{D}_{NT,r}^{2}\|_2 = O_P(1)$ and $\|\boldsymbol{D}_{NT,r}^{-2}\|_2 = O_P(1)$

B) For $i = 1, \ldots, N$ and $j = 1, \ldots, r$, we have for $q > 2$

$$P\left(\left|\sum_{s=1}^{T}\xi_{i,s}f_{j,s}\right| \geq x\right) \leq$$

$$\leq C_1\frac{T\max_{j,i}\|\{\xi_{j,t}f_{i,t}\}\|_{q,\alpha}}{x^q} + C_2\exp\left(-\frac{C_3x^2}{T\max_{j,i}\|\{\xi_{j,t}f_{i,t}\}\|_{2,\alpha}^2}\right).$$

Furthermore,

$$P\left(\max_{i,j}\left|\sum_{s=1}^{T}\xi_{i,s}f_{j,s}\right| \geq x\right) \leq$$
$$\leq C_1\frac{NT\max_{j,i}\|\{\xi_{j,t}f_{i,t}\}\|_{q,\alpha}}{x^q}+$$
$$+C_2\exp\left(-\frac{C_3x^2}{T\max_{j,i}\|\{\xi_{j,t}f_{i,t}\}\|_{2,\alpha}^2}+\log(N)\right).$$

This implies $\max_{i,j}|1/T\sum_{s=1}^{T}\xi_{i,s}f_{j,s}| = O_P(\sqrt{(\log(N)/T})+N^{2/\zeta}T^{2/\zeta-1})$.

C) For each $j=1,\ldots,r$*, we have for* $q>2$

$$P\left(\left|\sum_{s=1}^{T}(|f_{j,s}|^2-\Sigma_{F,j,j})\right| \geq x\right) \leq C_1\frac{T\max_i\|\{f_{i,t}^2\}\|_{q,\alpha}}{x^q}+$$
$$+C_2\exp\left(-\frac{C_3x^2}{T\max_i\|\{f_{i,t}^2\}\|_{2,\alpha}^2}\right).$$

Since r *is fixed, this implies* $1/T\sum_{s=1}^{T}|f_{s,j}|^2=\Sigma_{F,j,j}+O_P(1/\sqrt{T})$ *and* $\max_j|1/T\sum_{s=1}^{T}f_{s,j}\hat{f}_{s,l}| \leq (M+O_P(1/\sqrt{T}))^{1/2}$.

D) For each $j_1,j_2=1,\ldots,N$*, we have for* $q>2$

$$P\left(\left|\sum_{s=1}^{T}(\xi_{j_1,s}\xi_{j_2,s}-e_{j_1}^{\top}\Gamma_{\xi}e_{j_2})\right| \geq x\right) \leq C_1\frac{T\max_i\|\{\xi_{i,t}^2\}\|_{q,\alpha}}{x^q}+$$
$$+C_2\exp\left(-\frac{C_3x^2}{T\max_i\|\{\xi_{i,t}^2\}\|_{2,\alpha}^2}\right).$$

Furthermore,

$$P\left(\max_{j_1,j_2}\left|\sum_{s=1}^{T}(\xi_{j_1,s}\xi_{j_2,s}-e_{j_1}^{\top}\Gamma_{\xi}e_{j_2})\right| \geq x\right) \leq C_1\frac{NT\max_i\|\{\xi_{i,t}^2\}\|_{q,\alpha}}{x^q}+$$

$$+ C_2 \exp\left(-\frac{C_3 x^2}{T \max_i \|\{\xi_{i,t}^2\}\|_{2,\alpha}^2} + \log(N)\right),$$

which implies $\max_{j_1,j_2} |1/T \sum_{s=1}^T \xi_{s,j_1}\xi_{s,j_2}| \leq M + O_P(\sqrt{(\log(N)/T)} + N^{2/\zeta}T^{2/\zeta-1})$.

E) *For each* $k = 1, \ldots, r$ *we have* $e_k^\top \boldsymbol{\Lambda}^\top \Gamma_\xi(0) \boldsymbol{\Lambda} e_k / N \leq M^4/(1-\rho^2) < \infty$ *and*

$$P\left(\frac{1}{T}\sum_{s=1}^T \left(1/\sqrt{N}\sum_{i=1}^N \ell_{i,k}\xi_{i,s}\right)^2 \geq x\right) \leq C_1 \frac{T \max_i \|\{\xi_{i,t}^2\}\|_{q,\alpha}}{[(x - M^4/(1-\rho^2))T]^q} +$$
$$+ C_2 \exp\left(-\frac{C_3 (x - M^4/(1-\rho^2))^2}{\max_i \|\{\xi_{i,t}^2\}\|_{2,\alpha}^2}\right),$$

which implies $1/T \sum_{s=1}^T (1/\sqrt{N}\sum_{i=1}^N \ell_{i,k}\xi_{i,s})^2 = O_P(1)$.

F) *For* $q > 2$, *we have*

$$P\left(\max_{j,k}\left|\frac{1}{T}\sum_{s=1}^T e_j^\top \boldsymbol{\xi}_t \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda}^\top e_k - e_j^\top \Gamma_\xi(0) \boldsymbol{\Lambda}^\top e_k\right| \geq x\right) \leq C_1 \frac{NT \max_i \|\{\xi_{i,t}^2\}\|_{q,\alpha}}{x^q} +$$
$$+ C_2 \exp\left(-\frac{C_3 x^2}{T \max_i \|\{\xi_{i,t}^2\}\|_{2,\alpha}^2} + \log(N)\right).$$

Since $\|\Gamma_\xi(0)\boldsymbol{\Lambda}^\top\|_{\max} \leq \|\Gamma_\xi(0)\|_\infty \|\Lambda\|_{\max} \leq M^2 k_\xi$, *we have* $\max_{j,k} |1/T \sum_{s=1}^T e_j^\top \boldsymbol{\xi}_t \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda}^\top e_k| = O_P(k_\xi + \sqrt{\log(N)/T} + N^{2/\zeta}T^{2/\zeta-1})$.

G) *We have for* $j = 1, \ldots, N, l = 1, \ldots, r$

$$\left|\frac{1}{NT^2}\sum_{i=1}^N \sum_{s,t=1}^T \xi_{j,t} \boldsymbol{f}_t^\top \boldsymbol{\Lambda}_i \xi_{i,s} \hat{f}_{s,l}\right| \leq$$
$$\leq \left(\sum_{k=1}^r \left(\frac{1}{T}\sum_{t=1}^T \xi_{j,t} f_{k,t}\right)^2\right)^{1/2} \left(\sum_{k=1}^r \left(\frac{1}{N^2 T}\sum_{s=1}^T (e_k^\top \boldsymbol{\Lambda}^\top \boldsymbol{\xi}_s)^2\right)\left(\frac{1}{T}\sum_{s=1}^T \hat{f}_{l,s}^2\right)\right)^{1/2}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{NT}} \left(\sum_{k=1}^{r} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \xi_{j,t} f_{k,t} \right)^2 \right)^{1/2} \left(\sum_{k=1}^{r} \left(\frac{1}{NT} \sum_{s=1}^{T} (e_k^\top \boldsymbol{\Lambda}^\top \boldsymbol{\xi}_s)^2 \right) \right)^{1/2} \\
&= O_P\left(\frac{1}{\sqrt{NT}} \right),
\end{aligned}
$$

$$
\begin{aligned}
&\max_j |\tfrac{1}{NT^2} \textstyle\sum_{i=1}^{N} \sum_{s,t=1}^{T} \xi_{j,t} \boldsymbol{f}_t^\top \boldsymbol{\Lambda}_i \xi_{i,s} \hat{f}_{s,l}| = \\
&= O_P(\sqrt{\log(N)}/\sqrt{NT} + N^{2/\zeta-1/2} T^{2/\zeta-3/2}),
\end{aligned}
$$

$$
\begin{aligned}
&\max_j \left| \frac{1}{NT^2} \sum_{i=1}^{N} \sum_{s,t=1}^{T} \xi_{j,t} \xi_{i,t} \boldsymbol{\Lambda}_i \boldsymbol{f}_s \hat{f}_{l,s} \right| \leq \\
&\leq \left(\sum_{k=1}^{r} \max_j \left(\frac{1}{NT} \sum_{t=1}^{T} e_j^\top \boldsymbol{\xi}_t \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda}^\top e_k \right)^2 \right)^{1/2} \\
&\left(\sum_{k=1}^{r} \left(\frac{1}{T} \sum_{s=1}^{T} f_{k,s}^2 \right) \left(\frac{1}{T} \sum_{s=1}^{T} \hat{f}_{l,s}^2 \right) \right)^{1/2} \\
&\leq \frac{k_\xi}{N} \left(\sum_{k=1}^{r} \max_j \sum_{t=1}^{T} \left(\frac{1}{k_\xi T} e_j^\top \boldsymbol{\xi}_t \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda}^\top e_k \right)^2 \right)^{1/2} \left(\sum_{k=1}^{r} \left(\frac{1}{T} \sum_{s=1}^{T} f_{k,s}^2 \right) \right)^{1/2} \\
&= O_P\left(\frac{k_\xi}{N} + \frac{\sqrt{\log(N)}}{\sqrt{T}N} + N^{2/\zeta-1} T^{2/\zeta-1} \right),
\end{aligned}
$$

$$
\begin{aligned}
&\max_j \left| \frac{1}{NT^2} \sum_{i=1}^{N} \sum_{s,t=1}^{T} \xi_{j,t} \xi_{i,t} \xi_{i,s} f_{l,s} \right| = \\
&\quad = \max_j \left| \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{T} \xi_{j,t} \xi_{i,t} - e_j \Gamma_\xi(0) e_i + e_j \Gamma_\xi(0) e_i \right) \left(\frac{1}{T} \sum_{s=1}^{T} \xi_{i,s} f_{l,s} \right) \right| \\
&= \left\| \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{\xi}_t \boldsymbol{\xi}_t^\top - \Gamma_\xi(0) \right\|_{\max} \left\| \frac{1}{T} \sum_{s=1}^{T} \boldsymbol{\xi}_s f_{l,s} \right\|_{\max} + \\
&\quad + \left\| \frac{1}{T} \sum_{s=1}^{T} \boldsymbol{\xi}_s f_{l,s} \right\|_{\max} \left\| \Gamma_\xi(0) \right\|_\infty / N
\end{aligned}
$$

$$=O_P\left(\frac{\log(N)}{T}+(NT)^{4/\zeta}/T^2+\frac{k_\xi}{N}\left(\sqrt{(\log(N)/T)}+N^{2/\zeta}T^{2/\zeta-1}\right)\right)$$

and

$$\max_j\left|\frac{1}{NT^2}\sum_{i=1}^N\sum_{s,t=1}^T\xi_{j,t}\xi_{i,t}\xi_{i,s}\hat{f}_{l,s}\right|=$$
$$=O_P\left(\frac{\log(N)}{T}+N^{4/\zeta}T^{4/\zeta-2}+\frac{k_\xi}{N}+\frac{\sqrt{\log(N)}}{\sqrt{NT}}+N^{2/\zeta-1/2}T^{2/\zeta-1}\right).$$

This implies

$$\max_{j,l}\left|1/T\sum_{s=1}^T e_l^\top(\hat{\boldsymbol{f}}_s-H_{NT}\boldsymbol{f}_s)\xi_{i,s}\right|=$$
$$=O_P\Bigg(\frac{\log(N)}{T}+\frac{k_\xi}{N}+\frac{\sqrt{\log(N)}}{\sqrt{NT}}+$$
$$+(NT)^{2/\zeta}\left(\frac{1}{\sqrt{NT}}+\frac{1}{T^{3/2}}+(NT)^{2/\zeta}\frac{1}{T^2}\right)\Bigg).$$

H) We have for $t\in\mathbb{Z}$

$$\left|\frac{1}{NT}\sum_{i=1}^N\sum_{s=1}^T\xi_{i,t}\xi_{i,s}[\hat{\boldsymbol{f}}_s-H_{NT}\boldsymbol{f}_s]\right|\leq$$
$$\leq\frac{1}{N}\sum_{s=1}^N|\xi_{i,k}|\max_{j,l}\left|1/T\sum_{s=1}^T e_l^\top(\hat{\boldsymbol{f}}_s-H_{NT}\boldsymbol{f}_s)\xi_{i,s}\right|$$
$$=O_P\Bigg(\frac{\log(N)}{T}+\frac{k_\xi}{N}+\frac{\sqrt{\log(N)}}{\sqrt{NT}}+$$
$$+(NT)^{2/\zeta}\left(\frac{1}{\sqrt{NT}}+\frac{1}{T^{3/2}}+(NT)^{2/\zeta}\frac{1}{T^2}\right)\Bigg)$$

and

$$\left|\frac{1}{NT}\sum_{i=1}^N\sum_{s=1}^T e_l^\top\boldsymbol{f}_t^\top\boldsymbol{\Lambda}_i\xi_{i,s}[\hat{\boldsymbol{f}}_s-H_{NT}\boldsymbol{f}_s]\right|=$$

$$= O_P\bigg(\frac{\log(N)}{T} + \frac{k_\xi}{N} + \frac{\sqrt{\log(N)}}{\sqrt{NT}} + \\ + (NT)^{2/\zeta}\left(\frac{1}{\sqrt{NT}} + \frac{1}{T^{3/2}} + (NT)^{2/\zeta}\frac{1}{T^2}\right)\bigg).$$

This implies

$$\hat{\boldsymbol{f}}_{\boldsymbol{t}} - H_{NT}\boldsymbol{f}_{\boldsymbol{t}} = \\ = \frac{1}{NT}\left[\sum_{i=1}^{N}\sum_{s=1}^{T}\xi_{i,t}\boldsymbol{\Lambda}_i\boldsymbol{f}_s^\top\boldsymbol{H}_{NT}\boldsymbol{f}_s + \sum_{i=1}^{N}\sum_{s=1}^{T}\xi_{i,t}\xi_{i,s}\boldsymbol{H}_{NT}\boldsymbol{f}_s\right]\boldsymbol{D}_{NT,r}^{-2} \\ + O_P\bigg(\frac{\log(N)}{T} + \frac{k_\xi}{N} + \frac{\sqrt{\log(N)}}{\sqrt{NT}} + \\ + (NT)^{2/\zeta}\left(\frac{1}{\sqrt{NT}} + \frac{1}{T^{3/2}} + (NT)^{2/\zeta}\frac{1}{T^2}\right)\bigg).$$

I) We have

$$\max_{j,l}\left|\frac{1}{T}\sum_{s=1}^{T}f_{j,s}[\hat{\boldsymbol{f}}_s - H_{NT}\boldsymbol{f}_s]^\top e_l\right| = \\ = O_P\bigg(\frac{\log(N)}{T} + \frac{k_\xi}{N} + \frac{\sqrt{\log(N)}}{\sqrt{NT}} + \\ + (NT)^{2/\zeta}\left(\frac{1}{\sqrt{NT}} + \frac{1}{T^{3/2}} + (NT)^{2/\zeta}\frac{1}{T^2}\right)\bigg).$$

J) For each $j, l = 1, \ldots, r$

$$\frac{1}{T}\sum_{s=1}^{T}e_j^\top[\hat{\boldsymbol{f}}_s - H_{NT}\boldsymbol{f}_s][\hat{\boldsymbol{f}}_s - H_{NT}\boldsymbol{f}_s]^\top e_l \\ = O_P\bigg(\frac{\log(N)}{T} + \frac{k_\xi}{N} + \frac{\sqrt{\log(N)}}{\sqrt{NT}} + \\ + (NT)^{2/\zeta}\left(\frac{1}{\sqrt{NT}} + \frac{1}{T^{3/2}} + (NT)^{2/\zeta}\frac{1}{T^2}\right)\bigg).$$

K) We have

$$(\boldsymbol{H}_{NT}^\top)^{-1}\boldsymbol{\Lambda}_i - \hat{\boldsymbol{\Lambda}}_i = \frac{1}{T}\sum_{s=1}^{T} H_{NT}\boldsymbol{f}_s\xi_{i,s} + Error_i,$$

where

$$\max_i |Error_i| = O_P\Bigg(\frac{\log(N)}{T} + \frac{k_\xi}{N} + \frac{\sqrt{\log(N)}}{\sqrt{NT}} + \\ + (NT)^{2/\zeta}\left(\frac{1}{\sqrt{NT}} + \frac{1}{T^{3/2}} + (NT)^{2/\zeta}\frac{1}{T^2}\right)\Bigg)$$

Proof of Lemma 5.1. First note that under Assumption 10 and Remark 5.1 we have for $\alpha > 0.5, q \geq 4$ $\max_j \|\{\xi_{j,t}\}_t\|_{q,\alpha} < \infty, \max_j \|\{f_{i,t}\}_t\|_{q,\alpha} < \infty$ and $\max_{j,i} \|\{\xi_{j,t}f_{i,t}\}_t\|_{0.5q,\alpha} < \infty$. Furthermore, since $\{\boldsymbol{f}_t\}$ and $\{\boldsymbol{\xi}_t\}$ are linear processes and $\|B^{(j)}\|_2 \leq M\rho^j$, we have
$\max_{w\in\mathbb{R}^n, \|w\|_2\leq 1} \|\{w^\top\xi_t\}_t\|_{q,\alpha} < \infty$, $\max_j \|\{\xi_{j,t}^2\}\|_{q/2,\alpha} < \infty$ and
$\max_j \|\{f_{i,t}^2\}\|_{q/2,\alpha} < \infty$.

For part A, we have
$\boldsymbol{X}^\top\boldsymbol{X}/NT = \boldsymbol{\Lambda}\boldsymbol{F}^\top\boldsymbol{F}\boldsymbol{\Lambda}^\top/NT + \boldsymbol{\Xi}^\top\boldsymbol{\Xi}/NT + \boldsymbol{\Xi}^\top\boldsymbol{F}\boldsymbol{\Lambda}^\top/NT + \boldsymbol{\Lambda}\boldsymbol{F}^\top\boldsymbol{\Xi}/NT$.
Assumption 11, i.e, $\Sigma_\Lambda > 0, \Sigma_F > 0$, implies for N, T large enough that $\boldsymbol{F}/\sqrt{T}$ and $\boldsymbol{\Lambda}/\sqrt{N}$ have rank r and that all r eigenvalues are strictly positive. Furthermore, note that we have by part D and Assumption 9

$$\begin{aligned}\|\boldsymbol{\Xi}^\top\boldsymbol{\Xi}/NT\|_F^2 &= (1/N^2\sum_{i_1,i_2}^{N}(1/T\sum_{t=1}^{T}\xi_{i_1,t}\xi_{i_2,t})^2) = \\ &= (1/N^2\sum_{i_1,i_2}^{N}(e_{i_1}^\top\Gamma_\xi(0)e_{i_2} + 1/T\sum_{t=1}^{T}\xi_{i_1,t}\xi_{i_2,t} - e_{i_1}^\top\Gamma_\xi(0)e_{i_2})^2) = \\ &= O_P(k_\xi/N + \log(N)/T + (NT)^{1/\zeta}/T^2).\end{aligned}$$

Additionally, by part B we have

$$\|\boldsymbol{\Lambda}\boldsymbol{F}^\top\boldsymbol{\Xi}/NT\|_F^2 = \|\boldsymbol{\Xi}^\top\boldsymbol{F}\boldsymbol{\Lambda}^\top/NT\|_F^2 =$$

$$
\begin{aligned}
&= 1/N^2 \sum_{i_1,i_2}^{N} \sum_{l=1}^{r} (1/T \sum_{t=1}^{T} \xi_{i_1,t} f_{r,t})^2 \ell_{i,l} = \\
&= O_P(\log(N)/T + (NT)^{1/\zeta}/T^2).
\end{aligned}
$$

That means for N, T large the eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}/NT$ are approximately those of $\boldsymbol{\Lambda} \boldsymbol{F}^\top \boldsymbol{F} \boldsymbol{\Lambda}^\top/NT$. Hence, for N, T large, $\boldsymbol{X}^\top \boldsymbol{X}/NT$ possesses r positive eigenvalues which implies that $\boldsymbol{D}^2_{NT,r}$ is invertible and consequently, $\boldsymbol{D}^{-2}_{NT,r} = O_P(1)$. Since by Assumption 11 $\lim_T \|\boldsymbol{F}/\sqrt{T}\|_2 \leq M$ and $\lim_T \|\boldsymbol{\Lambda}/\sqrt{N}\|_2 \leq M$, we also have $\boldsymbol{D}^2_{NT,r} = O_P(1)$.

For the part B, note first that $\mathbb{E}\frac{1}{T}\sum_{s=1}^T \xi_{i,s} f_{j,s} = 0$ due to Assumption 10. Furthermore, since

$$
P(\max_{i,j} |\sum_{s=1}^T \xi_{i,s} f_{j,s}| \geq x) \geq \sum_{i,j} P(|\sum_{s=1}^T \xi_{i,s} f_{j,s}| \geq x),
$$

the assertion follows by Assumption 10 and Theorem 2 in Wu and Wu (2016). Since $\mathbb{E} f^2_{j,t} = \Sigma_{f,j,j}$ and $\mathbb{E}\xi_{j_1,t}\xi_{j_2 t} = e^\top_{j_1} \Gamma_\xi e_{j_2}$, Part C and D follow by the same arguments. Note also that for some vectors u, v and some symmetric matrix Γ, we have $u^\top \Gamma u \leq v^\top \Gamma v + u^\top \Gamma u$. That is why for $\max_{j_1,j_2} |\sum_{s=1}^T (\xi_{j_1,s}\xi_{j_2,s} - e^\top_{j_1} \Gamma_\xi e_{j_2})$ it is sufficient to look at $\max_j |\sum_{s=1}^T (\xi_{j,s}\xi_{j,s} - e^\top_j \Gamma_\xi e_j)$.

For the part E, note that $\Gamma_\xi(0) = \sum_{j=0}^\infty B^{(j)} \Sigma_v (B^{(j)})^\top$ and $1/\sqrt{N} \sum_{i=1}^N \ell_{i,k} \xi_{i,s} = 1/\sqrt{N} e^\top_k \Lambda \boldsymbol{\xi}_s$, where $1/\sqrt{N} e^\top_k \Lambda \in \mathbb{R}^N$ and $\|1/\sqrt{N} e^\top_k \Lambda\|_2 \leq M$. Since $\|B^{(j)}\|_2 \leq M\rho^j$ and $\|\Sigma_v\|_2 \leq M$ by Assumption 9,10, we have $\|\Gamma_\xi(0)\|_2 \leq M^3/(1-\rho^2)$ and the assertions follows then by Assumption 11 and part D. Part F follows by similar arguments.

The first four assertions in Part G follow by Cauchy-Schwartz and the previous parts of this lemma. For the fifth assertions note the following

$$
\frac{1}{NT^2} \sum_{i=1}^N \sum_{s,t=1}^T \xi_{j,t} \xi_{i,t} \xi_{i,s} \hat{f}_{l,s} =
$$

$$
\begin{aligned}
&= \sum_{k=1}^{r} \left(\frac{1}{NT^2} \sum_{i=1}^{N} \sum_{s,t=1}^{T} \xi_{j,t}\xi_{i,t}\xi_{i,s} f_{k,s} \right) \left(e_l^\top \boldsymbol{H}_{NT} e_k \right) + \\
&\frac{1}{NT^2} \sum_{i=1}^{N} \sum_{s,t=1}^{T} \xi_{j,t}\xi_{i,t}\xi_{i,s} (\hat{f}_{l,s} - e_l \boldsymbol{H}_{NT} \boldsymbol{f}_s) \\
&= I_j + II_j
\end{aligned}
$$

$\max_j |I_j| = O_P\left(\frac{\sqrt{\log(N)}}{T} + N^{2/\zeta} T^{2/\zeta - 3/2} + \frac{k_\xi}{N} \left(\sqrt{(\log(N)/T)} + N^{2/\zeta} T^{2/\zeta - 1} \right) \right)$. Furthermore, we have

$$
\begin{aligned}
&\max_j |II_j| = \max_j \left| \frac{1}{N^2T^3} \sum_{k=1}^{N} \sum_{s,t=1}^{T} \xi_{j,s}\xi_{k,s}\xi_{k,t} \left[\sum_{i=1}^{N} \sum_{s=1}^{T} \boldsymbol{f}_t^\top \boldsymbol{\Lambda}_i \xi_{i,s} \hat{\boldsymbol{f}}_s + \right. \right. \\
&\left. \left. \sum_{i=1}^{N} \sum_{s=1}^{T} \xi_{i,t} \boldsymbol{\Lambda}_i \boldsymbol{f}_s^\top \hat{\boldsymbol{f}}_s + \sum_{i=1}^{N} \sum_{s=1}^{T} \xi_{i,t}\xi_{i,s} \hat{\boldsymbol{f}}_s \right] \boldsymbol{D}_{NT,r}^{-2} e_l \right| \\
\leq & \frac{1}{\sqrt{N}} \max_{j,i} \left| \frac{1}{T} \sum_{s=1}^{T} \xi_{j,s}\xi_{i,s} \right| \max_{j,l} \left| \frac{1}{T} \sum_{t=1}^{T} \xi_{j,t} f_{l,t} \right| \left(\max_l \frac{1}{NT} \sum_{s=1}^{T} (e_l^\top \boldsymbol{\Lambda} \boldsymbol{\xi}_s)^2 \right)^{1/2} \quad \|\boldsymbol{D}_{NT}^{-2}\|_{\max} r^2 \\
+ & \max_{j,i} \left| \frac{1}{T} \sum_{s=1}^{T} \xi_{j,s}\xi_{i,s} \right| \max_{j,l} \left| \frac{1}{NT} \sum_{s=1}^{T} \xi_{j,t}\xi_t^\top \Lambda e_l \right| \max_{k,l} \left| \frac{1}{T} \sum_{s=1}^{T} f_{k,s}^\top \hat{f}_{l,s} \right| \|\boldsymbol{D}_{NT}^{-2}\|_{\max} r^2 \\
+ & \max_j \frac{1}{N^2} \left| e_j^\top \left(\frac{1}{T} \sum_{t=1}^{T} \xi_t \xi_t^\top \right) \left(\frac{1}{T} \sum_{t=1}^{T} \xi_t \xi_t^\top \right) \left(\frac{1}{T} \sum_{t=1}^{T} \xi_t \xi_t^\top \right) \right. \\
& \left. \left(\frac{1}{T} \sum_{t=1}^{T} \xi_t \xi_t^\top \right) \left(\frac{1}{T} \sum_{t=1}^{T} \xi_t \xi_t^\top \right) e_j \right|^{1/2} \\
= & O_P \left(1 + \left(\sqrt{\log(N)/T} + N^{2/\zeta} T^{2/\zeta - 1} \right) / \sqrt{N} \left(\sqrt{\log(N)/T} + N^{2/\zeta} T^{2/\zeta - 1} \right) \right) \\
+ & O_P \left(1 + \left(\sqrt{\log(N)/T} + N^{2/\zeta} T^{2/\zeta - 1} \right) \left(\frac{k_\xi}{N} + \frac{\sqrt{\log(N)}}{\sqrt{T}N} + N^{2/\zeta - 1} T^{2/\zeta - 1} \right) \right) + III \\
= & O_P \left(\frac{\log(N)}{T} + N^{4/\zeta} T^{4/\zeta - 2} + \frac{k_\xi}{N} + \frac{\sqrt{\log(N)}}{\sqrt{NT}} + N^{2/\zeta - 1/2} T^{2/\zeta - 1} \right)
\end{aligned}
$$

where

$$
III \leq 1/N^2 (\|(\frac{1}{T} \sum_{t=1}^{T} \xi_t \xi_t^\top) - \Gamma_\xi(0)\|_\infty + \|\Gamma_\xi(0)\|_\infty)^2 \| \frac{1}{T} \sum_{t=1}^{T} \xi_t \xi_t^\top \|_{\max}^{1/2} \leq
$$

$$
\begin{aligned}
&\leq (\|(\frac{1}{T}\sum_{t=1}^{T}\xi_t\xi_t^\top) - \Gamma_\xi(0)\|_{\max}^2 + k_\xi/N)^2\|\frac{1}{T}\sum_{t=1}^{T}\xi_t\xi_t^\top\|_{\max}^{1/2} = \\
&= O_P((\frac{\sqrt{\log(N)}}{\sqrt{T}} + N^{2/\zeta}T^{2/\zeta-1} + k_\xi/N)^2(1 + (\sqrt{\log(N)/T} + N^{2/\zeta}T^{2/\zeta-1}))).
\end{aligned}
$$

The sixth assertion is the combination of the previous assertions.

The first assertion in part H follows directly from part G. The second assertion follows from part B and Assumption 11. The third assertion follows by the same arguments as in part G. Note that in each assertion in part G $\xi_{j,t}$ is replaced with $\|e_l^\top \boldsymbol{\Lambda}^\top\|_2/N(e_l^\top \boldsymbol{\Lambda}^\top\|e_l^\top \boldsymbol{\Lambda}^\top\|_2\xi_t)$. Since $\|e_l^\top \boldsymbol{\Lambda}^\top\|_2/N = O(1/\sqrt{N})$ the assertion follows by part B,E.

For part I, we have for $j, l = 1, \ldots, r$

$$
\begin{aligned}
&\left|\frac{1}{T}\sum_{s=1}^{T} f_{j,s}[\hat{\boldsymbol{f}}_s - H_{NT}\boldsymbol{f}_s]^\top D_{NT,r}^2 e_l\right| = \\
=&\frac{1}{NT^2}\left|\sum_{t,s=1}^{T}\sum_{i=1}^{N} f_{j,t}\xi_{i,t}\boldsymbol{\Lambda}_i\boldsymbol{f}_s^\top \hat{f}_{s,l}\sum_{t,s=1}^{T} f_{j,t}\boldsymbol{\xi}_t^\top\boldsymbol{\xi}_s f_{s,l}\right| \\
+&O_P\left(\frac{\log(N)}{T} + \frac{k_\xi}{N} + \frac{\sqrt{\log(N)}}{\sqrt{NT}} + (NT)^{2/\zeta}\left(\frac{1}{\sqrt{NT}} + \frac{1}{T^{3/2}} + (NT)^{2/\zeta}\frac{1}{T^2}\right)\right) \\
\leq&\left(\sum_{k=1}^{r}\left(\frac{1}{TN}\sum_{t=1}^{T}(e_k^\top\boldsymbol{\Lambda}^\top\boldsymbol{\xi}_t f_{j,t})^2\right)^{1/2}\left(\sum_{k=1}^{r}\left(\frac{1}{T}\sum_{s=1}^{T} f_{k,r}^2\right)^{1/2}\right)\right) \\
+&\left[\left(\frac{1}{T}\sum_{t=1}^{T} f_{j,t}\boldsymbol{\xi}_t^\top\right)\left(\frac{1}{T}\sum_{s=1}^{T}\boldsymbol{\xi}_s\xi_s^\top\right)\left(\frac{1}{T}\sum_{z=1}^{T}\boldsymbol{\xi}_z f_{j,z}\right)\right]^{1/2}\frac{1}{N} \\
+&O_P\left(\frac{\log(N)}{T} + \frac{k_\xi}{N} + \frac{\sqrt{\log(N)}}{\sqrt{NT}} + (NT)^{2/\zeta}\left(\frac{1}{\sqrt{NT}} + \frac{1}{T^{3/2}} + (NT)^{2/\zeta}\frac{1}{T^2}\right)\right) \\
=&O_P\left(\frac{\log(N)}{T} + \frac{k_\xi}{N} + \frac{\sqrt{\log(N)}}{\sqrt{NT}} + (NT)^{2/\zeta}\left(\frac{1}{\sqrt{NT}} + \frac{1}{T^{3/2}} + (NT)^{2/\zeta}\frac{1}{T^2}\right)\right).
\end{aligned}
$$

.

Part J follows by part B,E, and H. For Part K, note that $\|\Lambda\|_{\max} \leq M$. Then, this part follows by G,I, and J. □

Proof of Theorem 5.1. (5.7) and (5.8) follow by Lemma 5.1. Furthermore, $w_{i,t} = \boldsymbol{\Lambda}_i^\top \boldsymbol{f}_t - \hat{\boldsymbol{\Lambda}}_i^\top \hat{\boldsymbol{f}}_t = \boldsymbol{\Lambda}_i^\top \boldsymbol{H}_{NT}^{-1}[\boldsymbol{H}_{NT}\boldsymbol{f}_t - \hat{\boldsymbol{f}}_t] + [(\boldsymbol{H}_{NT}^\top)^{-1}\boldsymbol{\Lambda}_i - \hat{\boldsymbol{\Lambda}}_i]^\top \boldsymbol{H}_{NT}\boldsymbol{f}_t + [(\boldsymbol{H}_{NT}^\top)^{-1}\boldsymbol{\Lambda}_i - \hat{\boldsymbol{\Lambda}}_i]^\top[\boldsymbol{H}_{NT}\boldsymbol{f}_t - \hat{\boldsymbol{f}}_t]$ and (5.9) follows by Lemma 5.1. □

Proof of Corollary 5.1. First note that under these assumptions, we have

$$w_{j,t} = \boldsymbol{\Lambda}_j^\top \boldsymbol{H}_{NT}^{-1}\frac{1}{NT}\left[\sum_{i=1}^{N}\sum_{s=1}^{T}\xi_{i,t}\boldsymbol{\Lambda}_i\boldsymbol{f}_s^\top \boldsymbol{H}_{NT}\boldsymbol{f}_s + \sum_{i=1}^{N}\sum_{s=1}^{T}\xi_{i,t}\xi_{i,s}\boldsymbol{H}_{NT}\boldsymbol{f}_s\right]\boldsymbol{D}_{NT,r}^{-2} +$$

$$+ \boldsymbol{f}_t^\top \boldsymbol{H}_{NT}^\top \frac{1}{T}\left[\sum_{s=1}^{T} H_{NT}\boldsymbol{f}_s\xi_{j,s}\right] + Error_j,$$

where $\max_i |Error_i| = O_P\Bigg(\frac{\log(N)}{T} +$

$+ \frac{k_\xi}{N} + \frac{\sqrt{\log(N)}}{\sqrt{NT}} + (NT)^{2/\zeta}\left(\frac{1}{\sqrt{N}T} + \frac{1}{T^{3/2}} + (NT)^{2/\zeta}\frac{1}{T^2}\right)\Bigg).$

For the first assertion, we have by Lemma 5.1

$$\Big\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_t\boldsymbol{\xi}_t^\top\Big\|_{\max} \leq \|\boldsymbol{\Lambda}\|_{\max}\|\boldsymbol{H}_{NT}\|_{\max}^2\|\boldsymbol{D}_{NT,r}^{-2}\|_{\max}$$

$$\Bigg[\Big\|\frac{1}{NT}\sum_{t=1}^{T}\boldsymbol{\Lambda}^\top\boldsymbol{\xi}_t\boldsymbol{\xi}_t^\top\Big\|_{\max}\Big\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_t^\top\boldsymbol{f}_t\Big\|_{\max} +$$

$$+ \max_i|Error_i| + \Big\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_s^\top\boldsymbol{\xi}_s\Big\|_{\max}\left(\Big\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t\boldsymbol{\xi}_t^\top - \Gamma_\xi(0)\Big\|_{\max} + \|\Gamma_\xi(0)\|_\infty/N\right)\Bigg] +$$

$$+ \Big\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_t^\top\boldsymbol{\xi}_t\Big\|_{\max}^2 = O_P\Bigg(\frac{k_\xi}{N} + \frac{\sqrt{\log(N)}}{N\sqrt{T}} + (NT)^{2/\zeta-1} + \frac{\log(N)}{T} + (NT)^{4/\zeta}/T^2 +$$

$$+ \frac{k_\xi}{N}\left(\sqrt{(\log(N)/T)} + N^{2/\zeta}T^{2/\zeta-1}\right)\Bigg) + O_P\Bigg(\frac{\log(N)}{T} + \frac{k_\xi}{N} + \frac{\sqrt{\log(N)}}{\sqrt{NT}} +$$

$$+ (NT)^{2/\zeta}\left(\frac{1}{\sqrt{N}T} + \frac{1}{T^{3/2}} + (NT)^{2/\zeta}\frac{1}{T^2}\right)\Bigg).$$

For the second assertion, we have by Lemma 5.1

$$
\begin{aligned}
&\Big\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_t\boldsymbol{w}_t^\top\Big\|_{\max} \leq \\
&\leq \|\boldsymbol{\Lambda}\|_{\max}^2\|\boldsymbol{H}_{NT}\|_{\max}^4\|\boldsymbol{D}_{NT,r}^{-2}\|_{\max}^2\Bigg[\Big\|\frac{1}{T}\sum_{s=1}^{T}f_s^\top f_s\Big\|_{\max}^2\Big\|\frac{1}{NT}\sum_{t=1}^{T}\boldsymbol{\Lambda}^\top\xi_t\boldsymbol{\xi}_t^\top\boldsymbol{\Lambda}\Big\|_{\max}/N \\
+&\Big\|\frac{1}{T}\sum_{s=1}^{T}f_s^\top f_s\Big\|_{\max}\Big\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_s^\top\boldsymbol{\xi}_s\Big\|_{\max}\left(2\Big\|\frac{1}{NT}\sum_{t=1}^{T}\boldsymbol{\Lambda}^\top\boldsymbol{\xi}_t\boldsymbol{\xi}_t\Big\|_{\max}+\Big\|\frac{1}{NT}\sum_{t=1}^{T}\boldsymbol{\Lambda}^\top\boldsymbol{\xi}_t f_t\Big\|_{\max}\right) \\
+&\Big\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_s^\top\boldsymbol{\xi}_s\Big\|_{\max}^2\left(\Big\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t\boldsymbol{\xi}_t^\top-\Gamma_\xi(0)\Big\|_{\max}+\|\Gamma_\xi(0)\|_\infty/N\right)+ \\
&+\Big\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_s^\top\boldsymbol{\xi}_s\Big\|_{\max}^3\Bigg]+\Big\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_s^\top\boldsymbol{\xi}_s\Big\|_{\max}^2\Big\|\frac{1}{T}\sum_{s=1}^{T}f_s^\top f_s\Big\|_{\max}+\max_i|Error_i| \\
=&O_P\left(\frac{1}{N}+\left(\frac{\sqrt{\log(N)}}{\sqrt{T}}+(NT)^{2/\zeta}/T\right)\left(\frac{k_\xi}{N}+\frac{\sqrt{\log(N)}}{\sqrt{T}}+(NT)^{2/\zeta}/T\right)\right) \\
+&O_P\left(\frac{\log(N)}{T}+\frac{k_\xi}{N}+\frac{\sqrt{\log(N)}}{\sqrt{NT}}+(NT)^{2/\zeta}\left(\frac{1}{\sqrt{NT}}+\frac{1}{T^{3/2}}+(NT)^{2/\zeta}\frac{1}{T^2}\right)\right) \\
=&O_P\left(\frac{k_\xi}{N}+\frac{\log(N)}{T}+\frac{\sqrt{\log(N)}}{\sqrt{NT}}+(NT)^{2/\zeta}\left(\frac{1}{\sqrt{NT}}+\frac{1}{T^{3/2}}+(NT)^{2/\zeta}\frac{1}{T^2}\right)\right).
\end{aligned}
$$

The third and fourth assertion follow by the same arguments. □

Lemma 5.2. *Under Assumption 9 and if*

$$
\Big\|\frac{1}{T-p}\sum_{t=p+1}^{T}(\hat{\xi}_{j,t}-\boldsymbol{\beta}^\top\hat{\boldsymbol{\xi}}_{t-1}^v)\hat{\boldsymbol{\xi}}_{t-1}^v\Big\|_{\max}\leq\hat{\lambda}/4
$$

and

$$
\Theta^\top\frac{1}{T-p}\sum_{t=p+1}^{T}\hat{\boldsymbol{\xi}}_{t-1}^v(\hat{\boldsymbol{\xi}}_{t-1,j})^\top\Theta\geq\alpha\|\Theta\|_2^2-\hat{\tau}\|\Theta\|_1^2\,\forall\,\Theta\in\mathbb{R}^{np}
$$

we have

$$
\|\hat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}^{(j)}\|_2\leq 16\max\left(\sqrt{k}(\hat{\lambda}/\alpha)^{1-q/2},\sqrt{\hat{\tau}}s(\hat{\lambda}/\alpha)\right)^{1-q}
$$

and

$\|\hat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}^{(j)}\|_1 \leq \max(68k(\hat{\lambda}/\alpha)^{1-q}, 64\sqrt{\tau}k^{3/2}(\hat{\lambda}/\alpha)^{1-3/2q} + 4k(\hat{\lambda}/\alpha)^{1-q})$.

Proof. This proof follows ideas of the Proof of Proposition 4.1 in Basu and Michailidis (2015b) as well as the Proof of Corollary 3 in Negahban et al. (2012). Let $\hat{\gamma} = 1/(T-p)\sum_{t=p+1}^{T} \hat{\xi}_{j,t}\hat{\xi}_{t-1}^{v}$ and $\hat{\Gamma} = 1/(T-p)\sum_{t=p+1}^{T} \hat{\xi}_{t-1}^{v}(\hat{\xi}_{t-1}^{v})^{\top}$. Let $\beta^* := \beta_j, \hat{\beta} := \hat{\beta}_j$ and $v = \hat{\beta} - \beta*$. Furthermore, let for some threshold $\eta > 0$ $J = J_\eta = \{j \in \{1, \ldots, np\} | e_j^{\top}\beta^*\| > \eta\}$ denote the set of indices for which β^* is absolutely greater than the threshold η, $\beta_n u$ refers to the hard thresholded vector with threshold η and for some vector u, u_J, u_{JC} denotes the vector obtained by the indices in J, J^C, respectively.

We have by Assumption 9 $\|\beta_\eta^* - \beta^*\|_1 \leq \eta^{1-q}k$. Furthermore, $|J| \leq \eta^{-q}k$.

Since $\hat{\beta}_j$ is the minimum given in (5.4), we have $-\hat{\beta}^{\top}\hat{\gamma} + \hat{\beta}^{\top}\hat{\Gamma}\hat{\beta} + \hat{\lambda}\|\hat{\beta}\|_1 \leq 2\beta^*\hat{\gamma} + (\beta^*)^{\top}\hat{\Gamma}\beta * + \hat{\lambda}\|\beta^*\|_1$. This gives further $v^{\top}\hat{\Gamma}v \leq 2v^{\top}(\hat{\gamma} - \hat{\Gamma}\beta*) + \hat{\lambda}(\|\beta * \|_1 - \|\beta^* + v\|_1) \leq 2v^{\top}(\hat{\gamma} - \hat{\Gamma}\beta*) + \hat{\lambda}(\|\beta_\eta^*\|_1 + 2\|\beta * - \beta_\eta^*\|_1 - \|\beta_\eta^* + v\|_1) \leq 2v^{\top}(\hat{\gamma} - \hat{\Gamma}\beta*) + \hat{\lambda}(\|v_J\|_1 - \|v_{JC}\|_1 + 2\eta^{1-q}k)$. This implies with the condition $\|\frac{1}{T-p}\sum_{t=p+1}^{T}(\hat{\xi}_{j,t} - \beta_j^{\top}\hat{\xi}_{t-1}^{v})\hat{\xi}_{t-1}^{v}\|_{\max} \leq \hat{\lambda}/4$ that $0 \leq v^{\top}\hat{\Gamma}v \leq 3/2\hat{\lambda}\|v_J\|_1 - 1/2\|v_{JC}\|_1 \leq 2\hat{\lambda}\|v\|_1 + 2\hat{\lambda}\eta^{1-q}k$. Hence, $\|v_{JC}\|_1 \leq 3\|v_J\|_1 + 4\eta^{1-q}k$ and since $|J| \leq \eta^{-q}k$, $\|v\|_1 \leq 4\sqrt{k}\eta^{-q/2}\|v\|_2 + 4s\eta^{1-q}$.

Then, with the condition $\Theta^{\top}\frac{1}{T-p}\sum_{t=p+1}^{T}\hat{\xi}_{t-1}^{v}(\hat{\xi}_{t-1,j})^{\top}\Theta \geq \alpha\|\Theta\|_2^2 - \hat{\tau}\|\Theta\|_1^2 \, \forall \, \Theta \in \mathbb{R}^{np}$ we obtain that $\alpha\|v\|_2^2 - \hat{\tau}\|v\|_1 \leq 8\hat{\lambda}\|v\|_2\sqrt{k}\eta^{-q/2} + 10\hat{\lambda}k\eta^{1-q}$. Set $\eta = \hat{\lambda}/\alpha$. Then, with the bound for $\|v\|_1$ and dropping minor terms in the maximum we obtain $\|v\|_2 \leq 16\max(\sqrt{k}(\hat{\lambda}/\alpha)^{1-q/2}, \sqrt{\hat{\tau}}s(\hat{\lambda}/\alpha)^{1-q})$. Furthermore,
$\|v\|_1 \leq \max(68k(\hat{\lambda}/\alpha)^{1-q}, 64\sqrt{\tau}k^{3/2}(\hat{\lambda}/\alpha)^{1-3/2q} + 4k(\hat{\lambda}/\alpha)^{1-q})$. □

Proof of Theorem 5.2. The idea is show determine the order of the quantities $\hat{\lambda}_T$ and $\hat{\tau}$ in Lemma 5.2. For this first note that since $(\xi_{j,t} - \beta_j^{\top}\boldsymbol{\xi}_{t-1}^{v})\boldsymbol{\xi}_{t-1}^{v} = v_{j,t}\boldsymbol{\xi}_{t_1}$ and $Ev_{j,t}\boldsymbol{\xi}_{t-1}^{v} = 0$, we have $\|\frac{1}{T-p}\sum_{t=p+1}^{T}(\boldsymbol{\xi}_{j,t} -$

$\beta_j^\top \boldsymbol{\xi}^v_{t-1})\boldsymbol{\xi}^v_{t-1}\|_{\max} = O_P(\sqrt{\log(Np)/T} + (NpT)^{2/\zeta}/T)$ by the same arguments as in the proof of Lemma 5.1 E and note that $\boldsymbol{\xi}^v_{t-1}$ is of dimension Np. Additionally, we have $\max_j \|\beta_j\|_1 \leq M^{1-q}k$ and

$$
\begin{aligned}
\hat{\lambda}_T = &\left\| \frac{1}{T-p} \sum_{t=p+1}^{T} (\hat{\xi}_{j,t} - \boldsymbol{\beta}^\top \hat{\xi}^v_{t-1})\hat{\boldsymbol{\xi}}^v_{t-1} \right\|_{\max} \leq \left\| \frac{1}{T-p} \sum_{t=p+1}^{T} (v_{j,t})\boldsymbol{\xi}^v_{t-1} \right\|_{\max} + \\
&+ \left\| \frac{1}{T-p} \sum_{t=p+1}^{T} v_{j,t}\boldsymbol{w}^v_{t-1} \right\|_{\max} \left\| \frac{1}{T-p} \sum_{t=p+1}^{T} w_{j,t}\boldsymbol{\xi}^v_{t-1} + w_{j,t}\boldsymbol{w}^v_{t-1} \right\|_{\max} + \\
&+ \|\boldsymbol{\beta}\|_1 \left\| \frac{1}{T-p} \sum_{t=p+1}^{T} \boldsymbol{w}^v_{t-1}(\boldsymbol{\xi}^v_{t-1})^\top + \boldsymbol{w}^v_{t-1}(\boldsymbol{w}^v_{t-1})^\top \right\|_{\max} \\
&= O_P\Bigg(\sqrt{\log(Np)/T} + (NpT)^{2/\zeta}/T + k\bigg(\frac{k_\xi}{N} + \frac{\log(Np)}{T} + \frac{\sqrt{\log(Np)}}{\sqrt{NT}} + \\
&+ (NpT)^{2/\zeta}\bigg(\frac{k_\xi}{NT} + \frac{1}{\sqrt{N}T} + \frac{1}{T^{3/2}} + (NpT)^{2/\zeta}\frac{1}{T^2} \bigg)\bigg)\Bigg).
\end{aligned}
$$

Let $\Gamma = \mathrm{Var}((\boldsymbol{\xi}_t^\top, \ldots, \boldsymbol{\xi}_{t-p+1}^\top)^\top)$ and $\hat{\Gamma} = \frac{1}{T-p}\sum_{t=p}^{T-1} \hat{\boldsymbol{\xi}}^v_t(\hat{\boldsymbol{\xi}}^v_t)^\top$. We have for

$$
\begin{aligned}
\Theta \in \mathbb{R}^{Np},\ \Theta^\top\hat{\Gamma}\Theta = \Theta^\top\Gamma\Theta + \Theta^\top(\hat{\Gamma} - \Gamma)\Theta \geq \alpha\|\Theta\|_2^2 - & \\
- \|\Theta\|_1^2\|\hat{\Gamma} - \Gamma\|_{\max} \geq \alpha\|\Theta\|_2^2 - \|\Theta\|_1^2\| \Bigg(& \|\Gamma - \\
- \frac{1}{T-p}\sum_{t=p}^{T-1} \boldsymbol{\xi}^v_t(\boldsymbol{\xi}^v_t)^\top\|_{\max} + 2\|\frac{1}{T-p}\sum_{t=p}^{T-1} \boldsymbol{\xi}^v_t(\boldsymbol{w}^v_t)^\top\|_{\max} + & \\
+ \frac{1}{T-p}\sum_{t=p}^{T-1} \hat{\boldsymbol{w}}^v_t(\boldsymbol{w}^v_t)^\top\|_{\max}\Bigg) =: \alpha\|\Theta\|_2^2 + \hat{\tau}\|\Theta\|_1^2. &
\end{aligned}
$$

With the results of Corollary 5.1 and Lemma 5.1, we have

$$
\begin{aligned}
\hat{\tau} = &O_P(\sqrt{\log(Np)/T} + (NpT)^{2/\zeta}/T + \\
&+ k\bigg(\frac{k_\xi}{N} + \frac{\log(Np)}{T} + \frac{\sqrt{\log(Np)}}{\sqrt{NT}} + (NpT)^{2/\zeta}\bigg(\frac{k_\xi}{NT} + \frac{1}{\sqrt{N}T} +
\end{aligned}
$$

$$+ \frac{1}{T^{3/2}} + (NpT)^{2/\zeta} \frac{1}{T^2} \Big) \Big).$$

That means $\hat{\tau} = O_p(\hat{\lambda}_T)$. Then, Lemma 5.2 give us

$$\begin{aligned} &\|\hat{\beta}_j - \beta_j\|_2 \leq 16 \max(\sqrt{k}(\hat{\lambda}_T/\alpha)^{1-q/2}, \sqrt{\hat{\tau}} k(\hat{\lambda}_T/\alpha)^{1-q} = \\ &= O_P(\sqrt{k}(\hat{\lambda}_T)^{1-q/2} + k(\hat{\lambda}_T)^{3/2-q/2}) \end{aligned}$$

and

$$\begin{aligned} &\|\hat{\beta}_j - \beta_j\|_1 \leq \max(68k(\hat{\lambda}_T/\alpha)^{1-q}, 64\sqrt{\tau}k^{3/2}(\hat{\lambda})^{1-3/2q} + \\ &+ 4k(\hat{\lambda}_T/\alpha)^{1-q}) = O_P(k(\hat{\lambda})^{1-q} + k^{3/2}(\hat{\lambda}_T)^{3/2(1-q)}) \\ &= O_P(k(\hat{\lambda})^{1-q}). \end{aligned}$$

□

Proof of Theorem 5.3. First note that

$$\begin{aligned} &e_j^\top(\hat{\boldsymbol{X}}_{T+1}^{1,p_2} - \boldsymbol{X}_{T+1}^{(1,p_2)}) = \hat{\boldsymbol{\beta}}_j \hat{\boldsymbol{\xi}}_T^v - \boldsymbol{\beta}_j \boldsymbol{\xi}_T^v + \hat{\boldsymbol{\Lambda}}_j^\top \sum_{i=1}^{p_2} \hat{\boldsymbol{\Pi}}_i^{(p_2)} \hat{\boldsymbol{f}}_{T+1-i} - \\ &- \boldsymbol{\Lambda}_j^\top \boldsymbol{H}_{NT}^{-1} \boldsymbol{H}_{NT} \sum_{i=1}^{p_2} \boldsymbol{\Pi}_i^{(p_2)} \boldsymbol{H}_{NT}^{-1} \boldsymbol{H}_{NT} \boldsymbol{f}_{T+1-i}. \end{aligned}$$

Then, the results derived in Theorem 5.1, 5.2 and Lemma 5.1 can be plugged in. Note further that due to due to Assumption 10 and 11 we have $\boldsymbol{\Lambda}\boldsymbol{\xi}_t/N = \|\boldsymbol{\Lambda}\|_2/N\boldsymbol{\Lambda}/\|\boldsymbol{\Lambda}\|_2\boldsymbol{\xi}_t = O_P(1/\sqrt{N})$ appearing in $\hat{\boldsymbol{f}}_{\boldsymbol{t}} - \boldsymbol{H}_{NT}\boldsymbol{f}_{\boldsymbol{t}}$ and the assertion follows. □

6 Conclusion

This concluding chapter discusses the main contributions of this thesis.

Vector autoregressive models (VARs) are the focus of this thesis in terms of time series modeling framework. These models are the workhorses for both estimating causes and effects in systems like the macroeconomy or the climate and to do forecasting. VARs very quickly become high-dimensional as the number of parameters to estimate increases quadratically with the number of time series included: an unrestricted VAR(p) has K^2p coefficients to be estimated, where K is the number of series and p is the lag-length. As the time series dimension T is typically fairly small for many economic and climate applications, the curse of dimensionality quickly affects standard least squares and maximum likelihood methods, making them unreliable. Therefore, VAR models are a natural framework for working with penalized regression techniques and factor models.

Within this framework, Chapter 2 is a necessary building block in this thesis as it extends the framework of “honest inference” in high-dimensional models to the stationary time series setting. Specifically, the post-double selection technique designed by Belloni, Chernozhukov, and Hansen (2014b) for treatment effect models, has been proven to be a fundamental tool to resolve the critique of Leeb and Pötscher (2005) and thus obtain uniformly valid inference also when the data are inherently time dependent. Chapter 2 is also of practical interest for what concerns the tuning of the penalty parameter in the lasso problem. When the aim of the analysis is inference, simulations reported in Chapter 2 distinguish the selection performances of various tuning techniques in terms of the size and power of the final test. Especially, information criteria such as the Bayesian Information Criterion (BIC) is shown to be a fast and easier solution compared to cross-validation and to often outperform it in practice.

Chapter 3 takes a step forward with respect to Chapter 2, as it allows for unit-root and cointegration among the time series in the VAR. "Allowing for unit root and cointegration" does not mean that one needs to test the integration order of the relevant time series; the true benefit of the procedure developed in Chapter 3 is in fact that one can entirely disregard whether unit roots and cointegration are present within the VAR. Classical unit root and cointegration tests depend on the exact model specification (intercept and/or deterministic time trend, the lag-length order, seasonality adjustments etc.) and they are therefore keen to biases and low statistical power. The methodology developed in Chapter 3 allows to skip completely the unit root and cointegration biased pretesting, thus also avoiding the practitioner to explicitly transform the series to stationary by taking their differences. The lag-augmentation proposed in Chapter 3 has also the great advantage of flexibility: the additional lags have the purpose of being used to internally take the differences of the integrated time series. However, no over-differencing occurs as the extra lags work "when needed": both time series integrated of order 0, 1, 2 are allowed as long as the lag-augmentation is at least 2 for all the variables interested in the hypothesis at test. As observed in the empirical application of Chapter 3 and in the climate econometrics setting of Chapter 4, the methodology developed in Chapter 3 is greatly beneficial to avoid bias and uncover causal relations. Concerning Chapter 4, climate time series are often found to be integrated of order 1 or

2 and having long memory. Avoiding to take first differences of the series renders the analysis more robust and allows to uncover causal connections between global temperature and other climate series which help in attributing climate change. The lag-augmentation idea also has some shortcomings that needs to be addressed. The whole argument of augmenting the lag-length essentially trades off statistical power of the performed test with bias otherwise incurred from pre-testing for unit-root and cointegration. In other words, it allows slightly less efficiency to gain in accurateness. The methodology developed in Chapter 3, and especially the possibility of only augmenting the lag-length of the variables of interest for the test, drastically reduces the power loss if compared to a full system augmentation of the lag-length. However, the power loss remains linked to the amount of variables involved in the hypothesis at test. If a full high-dimensional vector of parameters is the interest of the test, then this lag-augmentation idea fails as additional lags of all the series in the vector would need to be included letting the model become extremely high-dimensional and the statistical test extremely inefficient. However, as far as Granger causality is concerned, it is argued in Chapter 3 how usually one is interested, if not else for interpretability reasons, in testing bivariate relations conditional on a large information set. In such case the methodology of Chapter 3 works well in practice. Testing causality among blocks is therefore allowed within the framework of Chapter 3 although up to a certain degree of block dimensions.

The specific inferential focus of Chapters 2-4 is Granger causality, namely the interest is in performing hypothesis testing to assess questions of the type: “will the time series X_{t-1}, conditional on the information set Ω_{t-1} which contains all other available time series up to time $t-1$, be able to better predict Y_t than would Y_{t-1} itself?”. Working with high-dimensional data sets is a great opportunity from a causal analysis perspective. It allows to enlarge the information set with as many variables as practically allowed from the available set. The larger the information set, the less spurious causality is of concern.

Chapter 5 combines the positive aspects of both school of thoughts when it comes to dimensionality reduction: sparse versus dense. A dynamic factor model is shown to be able to be combined with a sparse VAR estimation of the idiosyncratic components. The benefit of the outlined two steps procedure in Chapter 5 is that it allows to disentangle in the system covariance matrix, the dependence among its diverging eigenvalues, namely the factors, with the dependence among the bounded ones i.e., the idiosyncratic components. Chapter 5 operates under the set up of a dynamic factor model with the aim of improving forecast accuracy. The common components are not assumed to be independent and identically distributed but are allowed to be dynamic. The type of dynamic considered however does not directly link with the original vector of time series thus making the relationship between them and the factors still static. Furthermore, the number of factors r is considered fixed as both the cross sectional dimension N and the time series dimension T grow large. This is argued to be a reasonable claim as assuming r to be a strictly increasing function in N or T would be tantamount to assume that all the eigenvalues of a large dimensional covariance matrix would necessarily diverge as the dimensions increase. Nonetheless, r needs to be estimated from the data and in Chapter 5 this issue is jointly considered with the one of lag-length estimation. A joint procedure to estimate the number of factors and the VAR lag-length is proposed, combining in an information criteria the approach of Bai and Ng (2002) to select the number of factor with an extra penalty allowing for simultaneous lag-length estimation.

In conclusion, this thesis extended honest inference to stationary and non-stationary high-dimensional time series models via the post-double selection technique. Sparsity-inducing techniques used to select the model were also integrated with dense dimensionality reduction techniques to obtain better forecast performances. Applications in both economics and climate science demonstrate how these methodologies can contribute to tackling important challenges in different fields.

Bibliography

Adamek, Robert, Stephan Smeekes, and Ines Wilms (2020). "LASSO inference for high-dimensional time series". In: *arXiv preprint arXiv: 2007.10952*.

Akaike, Hirotugu (1974). "A new look at the statistical model identification". In: *IEEE transactions on automatic control* 19.6, pp. 716–723.

Alessi, Lucia, Matteo Barigozzi, and Marco Capasso (2010). "Improved penalization for determining the number of factors in approximate factor models". In: *Statistics & Probability Letters* 80.23-24, pp. 1806–1813.

Altissimo, Filippo, Antonio Bassanetti, Riccardo Cristadoro, Mario Forni, Marc Hallin, Marco Lippi, Lucrezia Reichlin, and Giovanni Veronese (2001). "EuroCOIN: a real time coincident indicator of the euro area business cycle". In: *Available at SSRN 296860.*

Aruoba, S Boragan, Francis X Diebold, M Ayhan Kose, and Marco E Terrones (2010). *Globalization, the business cycle, and macroeconomic monitoring.* Tech. rep. National Bureau of Economic Research.

Aruoba, S Borağan and Francis X Diebold (2010). "Real-time macroeconomic monitoring: Real activity, inflation, and interactions". In: *American Economic Review* 100.2, pp. 20–24.

Attanasio, A and U Triacca (2011). "Detecting human influence on climate using neural networks based Granger causality". In: *Theoretical and Applied Climatology* 103.1-2, pp. 103–107.

Attanasio, Alessandro, Antonello Pasini, and Umberto Triacca (2012). "A contribution to attribution of recent global warming by out-of-sample Granger causality analysis". In: *Atmospheric Science Letters* 13.1, pp. 67–72.

Audrino, Francesco and Lorenzo Camponovo (2018). "Oracle Properties, Bias Correction, and Bootstrap Inference for Adaptive Lasso for Time Series M-Estimators". In: *Journal of Time Series Analysis* 39.2, pp. 111–128.

Bai, Jushan (Jan. 2003). "Inferential Theory for Factor Models of Large Dimensions". In: *Econometrica* 71.1, pp. 135–171.

Bai, Jushan and Serena Ng (2002). "Determining the number of factors in approximate factor models". In: *Econometrica* 70.1, pp. 191–221.

— (2019). "Rank regularized estimation of approximate factor models". In: *Journal of Econometrics* 212.1, pp. 78–96.

— (2020). "Simpler Proofs for Approximate Factor Models of Large Dimensions". In: *arXiv preprint arXiv:2008.00254*.

Baker, Scott R, Nicholas Bloom, Steven J Davis, and Kyle J Kost (2019). *Policy News and Stock Market Volatility*. Working Paper 25720. National Bureau of Economic Research.

Bańbura, Marta, Domenico Giannone, and Lucrezia Reichlin (2010). "Large Bayesian vector auto regressions". In: *Journal of Applied Econometrics* 25.1, pp. 71–92.

Banerjee, Anindya, Massimiliano Marcellino, and Igor Masten (2005). "Leading indicators for euro-area inflation and GDP growth". In: *Oxford Bulletin of Economics and Statistics* 67, pp. 785–813.

Barigozzi, Matteo and Christian Brownlees (2019). "Nets: Network estimation for time series". In: *Journal of Applied Econometrics* 34.3, pp. 347–364.

Barigozzi, Matteo and Marc Hallin (2017). "A network analysis of the volatility of high dimensional financial series". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 66.3, pp. 581–605.

Basu, Sumanta, Xianqi Li, and George Michailidis (2019). "Low rank and structured modeling of high-dimensional vector autoregressions". In: *IEEE Transactions on Signal Processing* 67.5, pp. 1207–1222.

Basu, Sumanta and George Michailidis (2015a). "Regularized estimation in sparse high-dimensional time series models". In: *Annals of Statistics* 43.4, pp. 1535–1567.

— (2015b). "Regularized estimation in sparse high-dimensional time series models". In: *The Annals of Statistics* 43.4, pp. 1535–1567.

Basu, Sumanta, Ali Shojaie, and George Michailidis (2015). "Network granger causality with inherent grouping structure". In: *The Journal of Machine Learning Research* 16.1, pp. 417–453.

Bauer, Gregory H and Keith Vorkink (2011). "Forecasting multivariate realized stock market volatility". In: *Journal of Econometrics* 160.1, pp. 93–101.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen (2012). "Sparse models and methods for optimal instruments with an application to eminent domain". In: *Econometrica* 80.6, pp. 2369–2429.

Belloni, Alexandre and Victor Chernozhukov (2013). "Least squares after model selection in high-dimensional sparse models". In: *Bernoulli* 19.2, pp. 521–547.

Belloni, Alexandre, Victor Chernozhukov, Ivan Fernández-Val, and Christian Hansen (2015). *Program evaluation with high-dimensional data.* Tech. rep. cemmap working paper.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2011a). "Inference for high-dimensional sparse econometric models". In: *arXiv preprint arXiv:1201.0220.*

— (2011b). "Lasso methods for gaussian instrumental variables models". In:

— (2014a). "High-dimensional methods and inference on structural and treatment effects". In: *Journal of Economic Perspectives* 28.2, pp. 29–50.

— (2014b). "Inference on treatment effects after selection among high-dimensional controls". In: *The Review of Economic Studies* 81.2, pp. 608–650.

Belloni, Alexandre, Victor Chernozhukov, and Kengo Kato (2014). "Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems". In: *Biometrika* 102.1, pp. 77–94.

Belloni, Alexandre, Victor Chernozhukov, and Lie Wang (2011). "Square-root lasso: pivotal recovery of sparse signals via conic programming". In: *Biometrika* 98.4, pp. 791–806.

— (2014). "Pivotal estimation via square-root lasso in nonparametric regression". In: *Annals of Statistics* 42.2, pp. 757–788.

Bennedsen, Mikkel, Eric Hillebrand, and Siem Jan Koopman (2020). "A statistical model of the global carbon budget". In: *EGU General Assembly Conference Abstracts*, p. 18986.

Bergmeir, Christoph, Rob J Hyndman, and Bonsoo Koo (2018). "A note on the validity of cross-validation for evaluating autoregressive time series prediction". In: *Computational Statistics & Data Analysis* 120, pp. 70–83.

Berk, Richard, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao (2013). "Valid post-selection inference". In: *Annals of Statistics* 41.2, pp. 802–837.

Bernanke, Ben S, Jean Boivin, and Piotr Eliasz (2005). "Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach". In: *The Quarterly journal of economics* 120.1, pp. 387–422.

Bickel, Peter J and Elizaveta Levina (2008). "Covariance regularization by thresholding". In: *The Annals of Statistics* 36.6, pp. 2577–2604.

Bickel, Peter J, Ya'acov Ritov, and Alexandre B Tsybakov (2009). "Simultaneous analysis of Lasso and Dantzig selector". In: *The Annals of statistics* 37.4, pp. 1705–1732.

Billio, Monica, Roberto Casarin, and Luca Rossini (2019). "Bayesian nonparametric sparse VAR models". In: *Journal of Econometrics* 212.1, pp. 97–115.

Billio, Monica, Mila Getmansky, Andrew W Lo, and Loriana Pelizzon (2012). "Econometric measures of connectedness and systemic risk in the finance and insurance sectors". In: *Journal of financial economics* 104.3, pp. 535–559.

Blanchard, Olivier J and Jordi Gali (2007). *The Macroeconomic Effects of Oil Shocks: Why are the 2000s so different from the 1970s?* Tech. rep. National Bureau of Economic Research.

Boivin, Jean and Serena Ng (2005). *Understanding and comparing factor-based forecasts.* Tech. rep. National Bureau of Economic Research.

— (2006). "Are more data always better for factor analysis?" In: *Journal of Econometrics* 132.1, pp. 169–194.

Bolt, Jutta and Jan Luiten van Zanden (2013). *The Maddison-Project.*

Brito, Diego, Marcelo C Medeiros, and Ruy Ribeiro (2018). *Forecasting large realized covariance matrices: The benefits of factor models and shrinkage.* Working Paper 3163668. SSRN.

Bühlmann, Peter and Sara Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media.

Cai, T Tony, Weidong Liu, and Harrison H Zhou (2016). "Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation". In: *The Annals of Statistics* 44.2, pp. 455–488.

Cai, T Tony, Zhao Ren, and Harrison H Zhou (2016). "Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation". In: *Electronic Journal of Statistics* 10.1, pp. 1–59.

Cai, Tony and Weidong Liu (2011). "Adaptive thresholding for sparse covariance matrix estimation". In: *Journal of the American Statistical Association* 106.494, pp. 672–684.

Callot, Laurent AF, Anders B Kock, and Marcelo C Medeiros (2017). "Modeling and forecasting large realized covariance matrices and portfolio choice". In: *Journal of Applied Econometrics* 32.1, pp. 140–158.

Cesa-Bianchi, Nicolo and Gábor Lugosi (2006). *Prediction, learning, and games.* Cambridge university press.

Chaudhry, Aditya, Pan Xu, and Quanquan Gu (2017). "Uncertainty assessment and false discovery rate control in high-dimensional Granger causal inference". In: *Proceedings of the 34th International Conference on Machine Learning.* Vol. 70, pp. 684–693.

Chen, Jiahua and Zehua Chen (2008). "Extended Bayesian information criteria for model selection with large model spaces". In: *Biometrika* 95.3, pp. 759–771.

— (2012). "Extended BIC for small-n-large-P sparse GLM". In: *Statistica Sinica*, pp. 555–574.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato (2013). "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors". In: *The Annals of Statistics* 41.6, pp. 2786–2819.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato (2014). “Gaussian approximation of suprema of empirical processes”. In: *Annals of Statistics* 42.4, pp. 1564–1597.

Chernozhukov, Victor, Chris Hansen, and Martin Spindler (2016). “High-dimensional metrics in R”. In: *arXiv preprint arXiv:1603.01700*.

Chernozhukov, Victor, Wolfgang K Härdle, Chen Huang, and Weining Wang (2020). “Lasso-driven inference in time and space”. In: *Annals of Statistics*, Forthcoming.

Chetverikov, Denis, Zhipeng Liao, and Victor Chernozhukov (2020). “On cross-validated lasso in high dimensions”. In: *Annal. Stat.* 40.

Choi, In (2012). “Efficient estimation of factor models”. In: *Econometric Theory*, pp. 274–308.

Chudik, Alexander and M Hashem Pesaran (2016). “Theory and practice of GVAR modelling”. In: *Journal of Economic Surveys* 30.1, pp. 165–197.

Church, John A, Neil J White, and Julie M Arblaster (2005). “Significant decadal-scale impact of volcanic eruptions on sea level and ocean heat content”. In: *Nature* 438.7064, pp. 74–77.

Clauset, Aaron, Mark EJ Newman, and Cristopher Moore (2004). “Finding community structure in very large networks”. In: *Physical review E* 70.6, p. 066111.

Cochrane, John H (1991). “A critique of the application of unit root tests”. In: *Journal of Economic Dynamics and Control* 15.2, pp. 275–284.

Corsi, Fulvio (2009). “A simple approximate long-memory model of realized volatility”. In: *Journal of Financial Econometrics* 7.2, pp. 174–196.

Corsi, Fulvio, Fabrizio Lillo, Davide Pirino, and Luca Trapin (2018). “Measuring the propagation of financial distress with Granger-causality tail risk networks”. In: *Journal of Financial Stability* 38, pp. 18–36.

Coulombe, Philippe Goulet and Maximilian Göbel (2021). *Arctic Amplification of Anthropogenic Forcing: A Vector Autoregressive Analysis.*

Coulombe, Philippe Goulet, Maxime Leroux, Dalibor Stevanovic, and Stéphane Surprenant (2020). "How is machine learning useful for macroeconomic forecasting?" In: *arXiv preprint arXiv:2008.12477*.

Cubadda, Gianluca, Alain Hecq, and Antonio Riccardo (2019). "Forecasting realized volatility measures with multivariate and univariate models". In: *Financial Mathematics, Volatility and Covariance Modelling*. Ed. by Julien Chevallier, Stéphane Goutte, David Guerreiro, Sophie Saglio, and Bilel Sanhaji. Vol. 2. Routledge. Chap. 11, pp. 286–308.

Davidson, James (1994). *Stochastic limit theory: An introduction for econometricians*. OUP Oxford.

Davis, Richard A, Pengfei Zang, and Tian Zheng (2016). "Sparse vector autoregressive modeling". In: *Journal of Computational and Graphical Statistics* 25.4, pp. 1077–1096.

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?" In: *Journal of Econometrics* 146, pp. 318–328.

Demirer, Mert, Francis X Diebold, Laura Liu, and Kamil Yilmaz (2018). "Estimating global bank network connectedness". In: *Journal of Applied Econometrics* 33.1, pp. 1–15.

Dickey, David A and Wayne A Fuller (1979). "Distribution of the estimators for autoregressive time series with a unit root". In: *Journal of the American statistical association* 74.366a, pp. 427–431.

Diebold, Francis X, Maximilian Göbel, Philippe Goulet Coulombe, Glenn D Rudebusch, and Boyuan Zhang (2020). "Optimal combination of Arctic sea ice extent measures: A dynamic factor modeling approach". In: *International Journal of Forecasting*.

Dlugokencky, E and Pieter Tans (2018). *Trends in atmospheric carbon dioxide, National Oceanic & Atmospheric Administration, Earth System Research Laboratory (NOAA/ESRL)*.

Doz, Catherine, Laurent Ferrara, and Pierre-Alain Pionnier (2020). "Business cycle dynamics after the great recession: an extended markov-switching dynamic factor model". In:

Dudley, Richard M (1978). “Central limit theorems for empirical measures”. In: *The Annals of Probability*, pp. 899–929.

Eichler, Michael (2013). “Causal inference with multiple time series: principles and problems”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371.1997, p. 20110613.

Engle, Robert (2002). “Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models”. In: *Journal of Business & Economic Statistics* 20.3, pp. 339–350.

Engle, Robert F (1984). “Wald, likelihood ratio, and Lagrange multiplier tests in econometrics”. In: *Handbook of econometrics* 2, pp. 775–826.

Estrada, Francisco, Dukpa Kim, and Pierre Perron (2021). “Anthropogenic influence in observed regional warming trends and the implied social time of emergence”. In: *Communications Earth & Environment* 2.1, pp. 1–9.

Fama, Eugene F and G William Schwert (1977). “Asset returns and inflation”. In: *Journal of financial economics* 5.2, pp. 115–146.

Fan, Jianqing, Yuan Ke, and Kaizheng Wang (2020). “Factor-adjusted regularized model selection”. In: *Journal of econometrics* 216.1, pp. 71–85.

Fan, Jianqing and Runze Li (2001). “Variable selection via nonconcave penalized likelihood and its oracle properties”. In: *Journal of the American statistical Association* 96.456, pp. 1348–1360.

Fan, Jianqing, Jinchi Lv, and Lei Qi (2011). “Sparse high-dimensional models in economics”. In: *Annual Review of Economics* 3.1, pp. 291–317.

Fan, Jianqing, Ricardo Masini, and Marcelo C Medeiros (2021). “Bridging factor and sparse models”. In: *arXiv preprint arXiv:2102.11341*.

Fava, Bruno and Hedibert F Lopes (2020). “The Illusion of the Illusion of Sparsity: An exercise in prior sensitivity”. In: *arXiv preprint arXiv:2009.14296*.

Forni, Mario and Luca Gambetti (2010). “The dynamic effects of monetary policy: A structural factor model approach”. In: *Journal of Monetary Economics* 57.2, pp. 203–216.

Forni, Mario, Alessandro Giovannelli, Marco Lippi, and Stefano Soccorsi (2018). “Dynamic factor model with infinite-dimensional factor space: Forecasting”. In: *Journal of Applied Econometrics* 33.5, pp. 625–642.

Forni, Mario, Marc Hallin, Marco Lippi, and Lucrezia Reichlin (2000). “The generalized dynamic-factor model: Identification and estimation”. In: *Review of Economics and statistics* 82.4, pp. 540–554.

— (2001). “Coincident and leading indicators for the euro area”. In: *The Economic Journal* 111.471, pp. C62–C85.

— (2003). “Do financial variables help forecasting inflation and real activity in the euro area?” In: *Journal of Monetary Economics* 50.6, pp. 1243–1255.

Forni, Mario, Marc Hallin, Marco Lippi, and Paolo Zaffaroni (2017). “Dynamic factor models with infinite-dimensional factor space: Asymptotic analysis”. In: *Journal of Econometrics* 199.1, pp. 74–92.

Friedlingstein, Pierre, Michael O’Sullivan, Matthew W Jones, Robbie M Andrew, Judith Hauck, Are Olsen, Glen P Peters, Wouter Peters, Julia Pongratz, and Stephen Sitch (2020). “Global carbon budget 2020”. In: *Earth System Science Data* 12.4, pp. 3269–3340.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2008). “Sparse inverse covariance estimation with the graphical lasso”. In: *Biostatistics* 9.3, pp. 432–441.

Friedrich, Marina, Stephan Smeekes, and Jean-Pierre Urbain (2020). “Autoregressive wild bootstrap inference for nonparametric trends”. In: *Journal of Econometrics* 214.1, pp. 81–109.

Friston, Karl, Rosalyn Moran, and Anil K Seth (2013). “Analysing connectivity with Granger causality and dynamic causal modelling”. In: *Current opinion in neurobiology* 23.2, pp. 172–178.

Fuller, Wayne A (2009). *Introduction to statistical time series*. Vol. 428. John Wiley & Sons.

Gao, Gelin, Bud Mishra, and Daniele Ramazzotti (2017). "Efficient Simulation of Financial Stress Testing Scenarios with Suppes-Bayes Causal Networks." In: *Procedia Computer Science* 108, pp. 272–284.

Geer, Sara van de, Peter Bühlmann, and Shuheng Zhou (2011). "The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso)". In: *Electronic Journal of Statistics* 5, pp. 688–749.

Geer, Sara van de and Johannes Lederer (2013). "The Bernstein–Orlicz norm and deviation inequalities". In: *Probability theory and related fields* 157.1-2, pp. 225–250.

Geer, Sara A van de (2016). *Estimation and testing under sparsity.* Springer.

Geweke, John (1977). "The dynamic factor analysis of economic time series". In: *Latent variables in socio-economic models.*

Giannone, Domenico, Michele Lenza, and Giorgio E Primiceri (2017). "Economic predictions with big data: The illusion of sparsity". In:

Giannone, Domenico, Lucrezia Reichlin, and David Small (2008). "Nowcasting: The real-time informational content of macroeconomic data". In: *Journal of Monetary Economics* 55.4, pp. 665–676.

Granger, Clive WJ (1969). "Investigating causal relations by econometric models and cross-spectral methods". In: *Econometrica: journal of the Econometric Society*, pp. 424–438.

— (1980). "Testing for causality: a personal viewpoint". In: *Journal of Economic Dynamics and control* 2, pp. 329–352.

Hallin, Marc and Roman Liška (2007). "Determining the number of factors in the general dynamic factor model". In: *Journal of the American Statistical Association* 102.478, pp. 603–617.

Hamilton, James Douglas (1994). *Time series analysis.* Vol. 2. Princeton university press Princeton, NJ.

Han, Fang, Huanran Lu, and Han Liu (2015). "A direct estimation of high dimensional stationary vector autoregressions". In: *The Journal of Machine Learning Research* 16.1, pp. 3115–3150.

Hansen, James, G Russell, A Lacis, I Fung, D Rind, and P Stone (1985). "Climate response times: Dependence on climate sensitivity and ocean mixing". In: *Science* 229.4716, pp. 857–859.

Hansen, James, Makiko Sato, Pushker Kharecha, Karina von Schuckmann, David J Beerling, Junji Cao, Shaun Marcott, Valerie Masson-Delmotte, Michael J Prather, and Eelco J Rohling (2017). "Young people's burden: requirement of negative CO 2 emissions". In: *Earth System Dynamics* 8.3, pp. 577–616.

Hansen, James E., Makiko Sato, Andrew Lacis, Reto Ruedy, Ina Tegen, and Elaine Matthews (1998). "Climate forcings in the Industrial era". In: *Proceedings of the National Academy of Sciences* 95.22, pp. 12753–12758. DOI: 10.1073/pnas.95.22.12753.

Hecq, Alain, Sébastien Laurent, and Franz C Palm (2016). "On the univariate representation of BEKK models with common factors". In: *Journal of Time Series Econometrics* 8.2, pp. 91–113.

Hiemstra, Craig and Jonathan D Jones (1994). "Testing for linear and nonlinear Granger causality in the stock price-volume relation". In: *The Journal of Finance* 49.5, pp. 1639–1664.

Hoerl, Arthur E and Robert W Kennard (1970). "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1, pp. 55–67.

Holton, James R, Peter H Haynes, Michael E McIntyre, Anne R Douglass, Richard B Rood, and Leonhard Pfister (1995). "Stratosphere-troposphere exchange". In: *Reviews of geophysics* 33.4, pp. 403–439.

Houghton, John Theodore, YDJG Ding, David J Griggs, Maria Noguer, Paul J van der Linden, Xiaosu Dai, Kathy Maskell, and CA Johnson (2001). *Climate change 2001: the scientific basis.* The Press Syndicate of the University of Cambridge.

Itō, Kiyosi and Kiyosi Itåo (2006). *Essentials of stochastic processes.* Vol. 231. American Mathematical Soc.

Javanmard, Adel and Andrea Montanari (2014). "Confidence intervals and hypothesis testing for high-dimensional regression". In: *The Journal of Machine Learning Research* 15.1, pp. 2869–2909.

Jiang, Wenxin (2009). "On uniform deviations of general empirical risks with unboundedness, dependence, and high dimensionality". In: *Journal of Machine Learning Research* 10.Apr, pp. 977–996.

Jing, Bing-Yi, Qi-Man Shao, and Qiying Wang (2003). "Self-normalized Cramér-type large deviations for independent random variables". In: *The Annals of probability* 31.4, pp. 2167–2215.

Johansen, Søren (1991). "Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models". In: *Econometrica: journal of the Econometric Society*, pp. 1551–1580.

— (1992). "A representation of vector autoregressive processes integrated of order 2". In: *Econometric theory* 8.2, pp. 188–202.

Joos, Fortunat and Renato Spahni (2008). "Rates of change in natural and anthropogenic radiative forcing over the past 20,000 years". In: *Proceedings of the National Academy of Sciences* 105.5, pp. 1425–1430.

Jordà, Òscar (2005). "Estimation and inference of impulse responses by local projections". In: *American economic review* 95.1, pp. 161–182.

Ju, Cheng, David Benkeser, and Mark J van Der Laan (2020). "Robust inference on the average treatment effect using the outcome highly adaptive lasso". In: *Biometrics* 76.1, pp. 109–118.

Kamińska, Anna, Han Ju Lee, and Hyung Joon Tag (2020). "Diameter two properties and the Radon-Nikodỳm property in Orlicz spaces". In: *Indagationes Mathematicae*.

Kaufmann, Robert and David Stern (1997). "Evidence for human influence on climate from hemispheric temperature relations". In: *Nature* 388, pp. 39–44.

Kim, Chang-Jin and Charles R Nelson (1998). "Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching". In: *Review of Economics and Statistics* 80.2, pp. 188–201.

Kiviet, Jan F (1986). "On the rigour of some misspecification tests for modelling dynamic relationships". In: *The Review of Economic Studies* 53.2, pp. 241–261.

Kneip, Alois and Pascal Sarda (2011). “Factor models and variable selection in high-dimensional regression analysis”. In: *Annals of statistics* 39.5, pp. 2410–2447.

Kock, Anders Bredahl (2016). “Consistent and conservative model selection with the adaptive lasso in stationary and nonstationary autoregressions”. In: *Econometric Theory* 32.1, p. 243.

Kock, Anders Bredahl and Laurent Callot (2015). “Oracle inequalities for high dimensional vector autoregressions”. In: *Journal of Econometrics* 186.2, pp. 325–344.

Kock, Anders Bredahl, Marcelo Medeiros, and Vasconcelos G (2020). “Penalized regressions”. In: *Macroeconomic Forecasting in the Era of Big Data.* Ed. by Peter Fuleky. Vol. 52. Advanced Studies in Theoretical and Applied Econometrics. Springer. Chap. 7, pp. 193–228.

Koopman, Siem Jan and Michel van der Wel (2013). “Forecasting the US term structure of interest rates using a macroeconomic smooth dynamic factor model”. In: *International Journal of Forecasting* 29.4, pp. 676–694.

Korobilis, Dimitris (2013). “Assessing the transmission of monetary policy using time-varying parameter dynamic factor models”. In: *Oxford Bulletin of Economics and Statistics* 75.2, pp. 157–179.

Korobilis, Dimitris and Davide Pettenuzzo (2019). “Adaptive hierarchical priors for high-dimensional vector autoregressions”. In: *Journal of Econometrics* 212.1, pp. 241–271.

Krampe, J, J-P Kreiss, and E Paparoditis (2018). *Bootstrap Based Inference for Sparse High-Dimensional Time Series Models.* arXiv e-print 1806.11083.

Krampe, Jonas and Efstathios Paparoditis (2021). “Sparsity Concepts and Estimation Procedures for High Dimensional Vector Autoregressive Models”. In: *Journal of Time Series Analysis.*

Lederer, Johannes and Michael Vogt (2020). “Estimating the Lasso’s Effective Noise”. In: *arXiv preprint arXiv:2004.11554.*

Lee, Jason D, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor (2016). “Exact post-selection inference, with application to the lasso”. In: *Annals of Statistics* 44.3, pp. 907–927.

Lee, Jim (2012). "Measuring business cycle comovements in Europe: Evidence from a dynamic factor model with time-varying parameters". In: *Economics Letters* 115.3, pp. 438–440.

— (2013). "Business cycle synchronization in Europe: Evidence from a dynamic factor model". In: *International Economic Journal* 27.3, pp. 347–364.

Leeb, Hannes and Benedikt M Pötscher (2005). "Model selection and inference: Facts and fiction". In: *Econometric Theory*, pp. 21–59.

Li, Kathleen T and David R Bell (2017). "Estimation of average treatment effects with panel data: Asymptotic theory and implementation". In: *Journal of Econometrics* 197.1, pp. 65–75.

Lin, Jiahe and George Michailidis (2017). "Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models". In: *The Journal of Machine Learning Research* 18.1, pp. 4188–4236.

— (2019). "Approximate Factor Models with Strongly Correlated Idiosyncratic Errors". In: *arXiv preprint arXiv:1912.04123.*

Liu, Weidong, Han Xiao, and Wei Biao Wu (2013). "Probability and moment inequalities under dependence". In: *Statistica sinica*, pp. 1257–1272.

Marcellino, Massimiliano, Mario Porqueddu, and Fabrizio Venditti (2016). "Short-term GDP forecasting with a mixed-frequency dynamic factor model with stochastic volatility". In: *Journal of Business & Economic Statistics* 34.1, pp. 118–127.

Martens, Martin (2004). *Estimating unbiased and precise realized covariances.* Working Paper 556118. SSRN.

Masini, Ricardo P, Marcelo C Medeiros, and Eduardo F Mendes (2019). "Regularized Estimation of High-Dimensional Vector AutoRegressions with Weakly Dependent Innovations". In: *arXiv preprint arXiv: 1912.09002.*

McAleer, Michael and Marcelo C Medeiros (2008). "Realized volatility: A review". In: *Econometric Reviews* 27.1-3, pp. 10–45.

McCracken, Michael W and Serena Ng (2016). "FRED-MD: A monthly database for macroeconomic research". In: *Journal of Business & Economic Statistics* 34.4, pp. 574–589.

Medeiros, Marcelo C and Eduardo F Mendes (2016a). "ℓ_1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors". In: *Journal of Econometrics* 191.1, pp. 255–271.

— (2016b). "ℓ_1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors". In: *Journal of Econometrics* 191.1, pp. 255–271.

Medeiros, Marcelo C, Gabriel FR Vasconcelos, Álvaro Veiga, and Eduardo Zilberman (2021). "Forecasting inflation in a data-rich environment: the benefits of machine learning methods". In: *Journal of Business & Economic Statistics* 39.1, pp. 98–119.

Meinshausen, Malte, Elisabeth Vogel, Alexander Nauels, Katja Lorbacher, Nicolai Meinshausen, David M Etheridge, Paul J Fraser, Stephen A Montzka, Peter J Rayner, and Cathy M Trudinger (2017). "Historical greenhouse gas concentrations for climate modelling". In: *Geoscientific Model Development* 10.5, pp. 2057–2116.

Meyer, Marco and Jens-Peter Kreiss (2015). "On the Vector Autoregressive Sieve Bootstrap". In: *Journal of Time Series Analysis* 36.3, pp. 377–397.

Miao, Ke, Peter CB Phillips, and Liangjun Su (2020). "High-Dimensional VARs with Common Factors". In:

Mitchell, John FB, TC Johns, Jonathan M Gregory, and SFB Tett (1995). "Climate response to increasing levels of greenhouse gases and sulphate aerosols". In: *Nature* 376.6540, pp. 501–504.

Morice, Colin P, John J Kennedy, Nick A Rayner, JP Winn, Emma Hogan, RE Killick, RJH Dunn, TJ Osborn, PD Jones, and IR Simpson (2020). "An updated assessment of near-surface temperature change from 1850: the HadCRUT5 dataset". In: *Journal of Geophysical Research: Atmospheres*, e2019JD032361.

Negahban, Sahand N, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu (2012). "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers". In: *Statistical science* 27.4, pp. 538–557.

Newman, Mark EJ and Michelle Girvan (2004). “Finding and evaluating community structure in networks”. In: *Physical Review E* 69.2, p. 026113.

Nicholson, William B, David S Matteson, and Jacob Bien (2017). “VARX-L: Structured regularization for large vector autoregressions with exogenous variables”. In: *International Journal of Forecasting* 33.3, pp. 627–651.

Nicholson, William B, Ines Wilms, Jacob Bien, and David S Matteson (2020). “High Dimensional Forecasting via Interpretable Vector Autoregression”. In: *Journal of Machine Learning Research* 21.166, pp. 1–52.

Oh, Dong Hwan and Andrew J Patton (2016). “High-dimensional copula-based distributions with mixed frequency data”. In: *Journal of Econometrics* 193.2, pp. 349–366.

Onatski, Alexei (2009). “A formal statistical test for the number of factors in the approximate factor models”. In: *Econometrica* 77.5, pp. 1447–1480.

Park, Joon Y and Peter CB Phillips (1988). “Statistical inference in regressions with integrated processes: Part 1”. In: *Econometric Theory*, pp. 468–497.

Pasini, Antonello, Umberto Triacca, and Alessandro Attanasio (2012). “Evidence of recent causal decoupling between solar radiation and global temperature”. In: *Environmental Research Letters* 7.3, p. 034020.

Phillips, Peter CB and Steven N Durlauf (1986). “Multiple time series regression with integrated processes”. In: *The Review of Economic Studies* 53.4, pp. 473–495.

Phillips, Peter CB and Victor Solo (1992). “Asymptotics for linear processes”. In: *The Annals of Statistics*, pp. 971–1001.

Plagborg-Møller, Mikkel and Christian K Wolf (2019). “Local projections and VARs estimate the same impulse responses”. In: *Unpublished paper: Department of Economics, Princeton University* 1.

Pooter, Michiel de, Martin Martens, and Dick van Dijk (2008). “Predicting the daily covariance matrix for S&P 100 stocks using intraday data—but which frequency to use?" In: *Econometric Reviews* 27.1-3, pp. 199–229.

Pretis, Felix (2020). "Econometric modelling of climate systems: The equivalence of energy balance models and cointegrated vector autoregressions". In: *Journal of Econometrics* 214.1, pp. 256–273.

Qiu, Weiliang and Harry Joe. (2020). *clusterGeneration: Random Cluster Generation (with Specified Degree of Separation)*. R package version 1.3.7.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ropelewski, Chester F and Phil D Jones (1987). "An extension of the Tahiti-Darwin southern oscillation index". In: *Monthly weather review* 115.9, pp. 2161–2165.

Rothenberg, TJ (1982). "Comparing alternative asymptotically equivalent tests". In: *Advances in econometrics*.

Sargent, Thomas J and Christopher A Sims (1977). "Business cycle modeling without pretending to have too much a priori economic theory". In: *New methods in business cycle research* 1, pp. 145–168.

Savin, N Eugene (1976). "Conflict among testing procedures in a linear regression model with autoregressive disturbances". In: *Econometrica: Journal of the Econometric Society*, pp. 1303–1315.

Schiavoni, Caterina, Franz Palm, Stephan Smeekes, and Jan van den Brakel (2021). "A dynamic factor model approach to incorporate Big Data in state space models for official statistics". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 184.1, pp. 324–353.

Schneider, Stephen H and Kristin Kuntz-Duriseti (2002). "Uncertainty and climate change policy". In: *Climate change policy: a survey*, pp. 53–87.

Schwarz, Gideon (1978). "Estimating the dimension of a model". In: *Annals of statistics* 6.2, pp. 461–464.

Seth, Anil K, Adam B Barrett, and Lionel Barnett (2015). "Granger causality analysis in neuroscience and neuroimaging". In: *Journal of Neuroscience* 35.8, pp. 3293–3297.

Sims, Christopher A (1972). “Money, income, and causality”. In: *The American economic review* 62.4, pp. 540–552.

Sims, Christopher A, James H Stock, and Mark W Watson (1990). “Inference in linear time series models with some unit roots”. In: *Econometrica: Journal of the Econometric Society*, pp. 113–144.

Skripnikov, A and G Michailidis (2019). “Joint estimation of multiple network Granger causal models”. In: *Econometrics and Statistics* 10, pp. 120–133.

Smeekes, Stephan and AM Robert Taylor (2012). “Bootstrap union tests for unit roots in the presence of nonstationary volatility”. In: *Econometric Theory*, pp. 422–456.

Smeekes, Stephan and Etienne Wijler (2018). “Macroeconomic forecasting using penalized regression methods”. In: *International journal of forecasting* 34.3, pp. 408–430.

— (2021). “An automated approach towards sparse single-equation cointegration modelling”. In: *Journal of Econometrics* 221.1, pp. 247–276.

Solomon, Susan, Martin Manning, Melinda Marquis, and Dahe Qin (2007). *Climate change 2007-the physical science basis: Working group I contribution to the fourth assessment report of the IPCC.* Vol. 4. Cambridge university press.

Song, Song and Peter J Bickel (2011). *Large vector auto regressions.* arXiv e-print 1106.3915.

Song, Xiaojun and Abderrahim Taamouti (2019). “A better understanding of granger causality analysis: A big data environment”. In: *Oxford Bulletin of Economics and Statistics* 81.4, pp. 911–936.

Stern, David and Robert Kaufmann (2014). “Anthropogenic and natural causes of climate change”. In: *Climatic Change* 122, pp. 257–269.

Stock, James H and Mark W Watson (1989a). “Interpreting the evidence on money-income causality”. In: *Journal of Econometrics* 40.1, pp. 161–181.

— (1989b). “New indexes of coincident and leading economic indicators”. In: *NBER macroeconomics annual* 4, pp. 351–394.

— (1999). "Forecasting inflation". In: *Journal of Monetary Economics* 44.2, pp. 293–335.

— (2002a). "Forecasting using principal components from a large number of predictors". In: *Journal of the American statistical association* 97.460, pp. 1167–1179.

— (2002b). "Macroeconomic forecasting using diffusion indexes". In: *Journal of Business & Economic Statistics* 20.2, pp. 147–162.

Stucky, Benjamin and Sara Van De Geer (2017). "Sharp oracle inequalities for square root regularization". In: *Journal of Machine Learning Research* 18.67, pp. 1–29.

Sun, De-Zheng and Kevin E Trenberth (1998). "Coordinated heat removal from the equatorial Pacific during the 1986–87 El Nino". In: *Geophysical research letters* 25.14, pp. 2659–2662.

Thompson, David WJ, John J Kennedy, John M Wallace, and Phil D Jones (2008). "A large discontinuity in the mid-twentieth century in observed global-mean surface temperature". In: *Nature* 453.7195, pp. 646–649.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Toda, Hiro Y and Taku Yamamoto (1995). "Statistical inference in vector autoregressions with possibly integrated processes". In: *Journal of econometrics* 66.1-2, pp. 225–250.

Trapani, Lorenzo (2018). "A randomized sequential procedure to determine the number of factors". In: *Journal of the American Statistical Association* 113.523, pp. 1341–1349.

Triacca, U (2001). "On the use of Granger causality to investigate the human influence on climate". In: *Theoretical and applied climatology* 69, pp. 137–138.

— (2005). "Is Granger causality analysis appropriate to investigate the relationship between atmospheric concentration of carbon dioxide and global surface air temperature?" In: *Theoretical and applied climatology* 81.3-4, pp. 133–135.

Triacca, Umberto, Alessandro Attanasio, and Antonello Pasini (2013). "Anthropogenic global warming hypothesis: testing its robustness by Granger causality analysis". In: *Environmetrics* 24.4, pp. 260–268.

Van de Geer, Sara, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure (2014). "On asymptotically optimal confidence regions and tests for high-dimensional models". In: *Annals of Statistics* 42.3, pp. 1166–1202.

Van Der Vaart, Aad W and Jon A Wellner (1996). "Weak convergence". In: *Weak convergence and empirical processes.* Springer, pp. 16–28.

Vyrost, Tomaš, Štefan Lyócsa, and Eduard Baumöhl (2015). "Granger causality stock market networks: Temporal proximity and preferential attachment". In: *Physica A: Statistical Mechanics and its Applications* 427, pp. 262–276.

Wainwright, Martin J (2019). *High-dimensional statistics: A non asymptotic viewpoint.* Vol. 48. Cambridge University Press.

Wang, Hansheng and Chenlei Leng (2007). "Unified LASSO estimation by least squares approximation". In: *Journal of the American Statistical Association* 102.479, pp. 1039–1048.

Wang, Hansheng, Bo Li, and Chenlei Leng (2009). "Shrinkage tuning parameter selection with a diverging number of parameters". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.3, pp. 671–683.

Wang, Hansheng, Runze Li, and Chih-Ling Tsai (2007). "Tuning parameter selectors for the smoothly clipped absolute deviation method". In: *Biometrika* 94.3, pp. 553–568.

Weiner, N (1956). *The Theory of Prediction'in EF Beckenback (ed) Modern Mathematics for Engineers.*

Wilms, Ines and Christophe Croux (2018). "An algorithm for the multivariate group lasso with covariance estimation". In: *Journal of Applied Statistics* 45.4, pp. 668–681.

Wilms, Ines, Sarah Gelper, and Christophe Croux (2016). "The predictive power of the business and bank sentiment of firms: A high-dimensional Granger Causality approach". In: *European Journal of Operational Research* 254.1, pp. 138–147.

Windmeijer, Frank, Helmut Farbmacher, Neil Davies, and George Davey Smith (2019). "On the use of the lasso for instrumental variables estimation with some invalid instruments". In: *Journal of the American Statistical Association* 114.527, pp. 1339–1350.

Wong, Kam Chung, Zifan Li, and Ambuj Tewari (2020). "Lasso guarantees for β-mixing heavy-tailed time series". In: *Annals of Statistics* 48.2, pp. 1124–1142.

Wooldridge, Jeffrey M (1987). "A regression based Lagrange Multiplier statistic that is robust in the presence of heteroskedasticity". In:

— (2015). *Introductory Econometrics: A Modern Approach.* Nelson Education.

Wu, Wei Biao (2005). "Nonlinear system theory: Another look at dependence". In: *Proceedings of the National Academy of Sciences* 102.40, pp. 14150–14154.

Wu, Wei-Biao and Ying Nian Wu (2016). "Performance bounds for parameter estimates of high-dimensional linear models with correlated errors". In: *Electronic Journal of Statistics* 10.1, pp. 352–379.

Yuan, Ming and Yi Lin (2006). "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1, pp. 49–67.

Zanna, Laure, Samar Khatiwala, Jonathan M Gregory, Jonathan Ison, and Patrick Heimbach (2019). "Global reconstruction of historical ocean heat storage and transport". In: *Proceedings of the National Academy of Sciences* 116.4, pp. 1126–1131.

Zellner, Arnold and Franz C Palm (1974). "Time series analysis and simultaneous equation econometric models". In: *Journal of Econometrics* 2.1, pp. 17–54.

Zhang, Cun-Hui and Stephanie S Zhang (2014). "Confidence intervals for low dimensional parameters in high dimensional linear models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1, pp. 217–242.

Zhang, Danna and Wei Biao Wu (2017). "Gaussian approximation for high dimensional time series". In: *The Annals of Statistics* 45.5, pp. 1895–1919.

Zhang, Rongmao, Peter Robinson, and Qiwei Yao (2019). "Identifying cointegration by eigenanalysis". In: *Journal of the American Statistical Association* 114.526, pp. 916–927.

Zhao, Peng and Bin Yu (2006). “On model selection consistency of Lasso”. In: *The Journal of Machine Learning Research* 7, pp. 2541–2563.

Zivot, Eric and Jiahui Wang (2003). “Long Memory Time Series Modeling”. In: *Modeling Financial Time Series with S-Plus®*. New York, NY: Springer New York, pp. 257–297. DOI: `10.1007/978-0-387-21763-5_8`.

Zou, Hui (2006). “The adaptive lasso and its oracle properties”. In: *Journal of the American statistical association* 101.476, pp. 1418–1429.

Zou, Hui and Trevor Hastie (2005). “Regularization and variable selection via the elastic net”. In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2, pp. 301–320.

Valorisation

The following chapter discusses the value of this dissertation for society.

As all the reported chapters of this dissertation deal with high-dimensional time series models, the importance of dealing with such models is now spelled out in the context of the current societal state.

Today's world evolves at fast pace. This rapid evolving state was reached thanks to the technological revolution started with the invention of the first computing machines, already before the twentieth century. Ever since the idea of modern computer coined by Alan Turing later in 1936, the history of technological advancements has grown beyond unimaginable limits. Very many pieces of technology became everyday objects for almost everyone in the world. Technology is most essentially a set of tools to facilitate complex tasks while drastically reducing the required time to complete those tasks. To a (very) simplified extent, one could say that technology equals "seeking (more) efficiency". But there is more to it than just speed. From a data perspective, the world soon started to realize how electronic machines are not only capable to rapidly compute tasks but also to store outputs. These outputs carry information and information is fundamental to optimize all kinds of human processes. Data then became literally an asset, even a currency, in the present state of the world. Although not explicitly used throughout this dissertation, the term "big data" is a popular broad expression to indicate the nowadays possibility of collecting data both at high-frequency and for very many different variables. The latter aspect of the large number of variables, often larger than the sample size, is what has been the focus of this dissertation and what has been labeled throughout as "high-dimensions". Many are the examples that can be given to describe this data abundance offered by technology. Mobile phones, out of all technological pieces, are probably the most fitting example. Close to everyone owns one and, if allowed, these pieces of hardware are capable to store various information about the owner, such as: location (through integrated GPS technology), habits, preferences, health information (see e.g., smartwatches linked to the phone) and many more. Without entering in the territory of online privacy which is besides the point to be made here, it is evident that gaining access to large quantity of very diverse data is nowadays a very easy task.

Beyond technological machines, the web has also played an important role in creating a need for high-dimensional techniques. While once finding data was costly, nowadays is often a couple of mouse clicks away. This allows for aggregating data sets containing data from multiple sources making them richer and, as a consequence, larger.

On a general standpoint, what is the value added of all this data? Everyone grew up thinking about mathematical, statistical, economic models as a great simplification of the reality but still useful enough to base some of our decisions on their results. This largely remains the case. However, the advent of the era of "big data" is one great opportunity to take a stand from oversimplified models and to approach a step higher in the ladder towards a better explanation of the complex reality. Being able to build models containing large amount of variables is tantamount to allow for richer information sets to condition the relations of interest upon, thus making the results more robust.

Nevertheless, just having large quantities of variables does not prove any useful if reliable techniques to handle and statistically analyze them are not developed. As explained in several parts of this dissertation, blessings and curses accompany the statistical treatment of large dimensional data sets. Techniques able to circumvent the curses while retaining as much as possible of the blessings are therefore paramount to navigate the high-dimensions.

This dissertation has focused on a specific data type to deal with in high-dimensions, namely time series. With respect to cross-sectional data, time series introduce an extra layer of complication which is the inherent time dependence. More specifically, techniques allowing for hypothesis testing in high-dimensional stationary and unit root non-stationary time series models have been developed in Chapter 2 and 3 and applied in Chapter 4. Techniques to enhance the forecast accuracy in high-dimensional models have been instead developed in Chapter 5.

Central to Chapter 2 and 3 is the question of causality. Establishing causes and effects among variables is clearly among the most basic yet most complicated tasks to face. The pioneering works of Clive Granger on causality have shown how data constraints require the causality concept to be reduced to an operational form. As textbooks state, correlation does not imply causation and unless one is able to condition the relation of interest on all the available information in the universe, then no true causality is possibly found. While this remains true, the high-dimensional framework precisely allows for much broader conditional sets of variables. This makes causal findings more robust than was ever possible before.

Causal questions are ubiquitous in probably all the fields of science. An application to finance is reported in Chapter 1 where stock realized volatilities are tested for pairwise Granger causality conditioning on all the other available stocks. The obtained networks of “spillover” effects are important tools to predict the flow of contagion when a financial crisis hits the market.

Chapter 4 is instead entirely dedicated to another, very relevant application of the high-dimensional Granger causality testing framework developed in Chapter 2, namely climate change. Climate econometrics is a sub-field of econometrics which arose in the last few years, in response to the urgent and pressing matter of climate change. For long time climate scientists have warned governments throughout the globe against the impact of climate change, but such warnings have long been ignored. Only in recent years some steps forward have been made in fixing targets to reduce CO_2 emission throughout the world. As the climate is a complex system, climate change research have seen different scientific fields joining forces in tackling different aspects of the problem. The framework of causality developed in Chapter 3 fits the purpose of climate attribution, namely advancing the understanding of the factors most responsible for the changing of global temperature. In Chapter 4 the high-dimensional causal framework allows to obtain clearer pictures of which climate variables are most affected by emissions through time and which causal relations exist among different anthropogenic emissions and temperature. These high-dimensional causal discoveries are relevant for policymakers to better understand the most pressing factors that need to be tackled in order to scale down the effects of the damage already done on the climate and avoid further damage.

In sum, the research presented in this dissertation is relevant for both academics and professionals across the fields of economics, finance, climate science and possibly other fields where investigating causes and effects, or obtain accurate forecasts, is relevant. In fact, the high-dimensional nature of the data available nowadays has been exploited within the presented methodologies and this has contributed in making findings more robust.

Curriculum Vitae

Luca Margaritella was born on April 29, 1993 in Milan, Italy. Between 2012 and 2015 he studied at the University of Milan-Bicocca where he obtained his bachelor degree (BSc) in Statistics. Between 2016 and 2017 he studied at Maastricht University, The Netherlands, obtaining a master degree (MSc) in Econometrics and Operation Research. In 2017 Luca started his PhD in Econometrics at Maastricht University under the supervision of Dr. Stephan Smeekes and Prof. dr. Alain Hecq. The findings of the research carried out from 2017 to 2021 are presented in this dissertation.

In September 2021 Luca started as assistant professor at Lund University, Sweden.