Three essays on time-varying parameters and time series networks

Rothfelder, Mario

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

THREE ESSAYS ON TIME-VARYING PARAMETERS AND TIME SERIES NETWORKS

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. E.H.L. Aarts, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op vrijdag 16 maart 2018 om 14.00 uur door

MARIO PHILIPP ROTHFELDER

geboren op 1 april 1987 te Mindelheim, Duitsland.

PROMOTIECOMMISSIE:

PROMOTOR:	prof. dr. B.J.M. Werker
COPROMOTOR:	dr. O. Boldea
OVERIGE LEDEN:	prof. dr. A. Lucas
	prof. dr. B. Melenberg
	dr. A. Pick
	dr. N.F.F. Schweizer

To Elli and Nicki Feger.

Acknowledgements

The past six years - or more precisely the last eleven years since I started my studies at the University of Konstanz - that led to this doctoral thesis have been an incredible journey for me, both personally and academically. I cannot and do not want to pretend that this endeavour would have been possible without the help, support and encouragement of many other extraordinary individuals whom I met along the way. Without them, this thesis would never have come together.

First, and foremost, I want to thank my co-promoter Otilia, under whose supervision this thesis took shape. I approached you in the last lecture of the course Econometrics 3 at the end of the first year of the Research Master to express my interest in a possible PhD project. Subsequently, you gave me an outstanding lead into time varying parameter problems, from which the second chapter of this thesis eventually originated. I am grateful for the freedoms you gave me in choosing topics for the other chapters of this thesis. As a supervisor you did not only put a tremendous amount of time and effort in my academic development but also supported me along the way when my self-confidence plumetted or other struggles arose. Your door was always open to seek advice or just have a talk.

I am also deeply grateful to my promoter Bas. You also always had an open door and I cannot thank you enough for your insightful comments and advice, even though I usually left our talks with even more unanswered questions than I initially had. I have learned a lot from you, not only during our talks but also by attending several of your lectures and your comments and remarks throughout the seminar series. I am more than happy to be able to call the two of you my “Doktormutter” and “Doktorvater”, and that our collaboration will continue after I finished my doctoral studies.

In addition, I want to thank my doctoral committee - André, Andreas, Bertrand and Nikolaus - for putting so much time and effort into reading the first draft of this thesis. I am indebted to all of you for providing me with wour comments, remarks and thought during the pre-defense. It is truely an honor to draw from your expertise.

During the last seven years in Tilburg I met many people which enriched my life and deserve some words of gratitude. First, my three office mates Alaa, Yeqiu and Daniël who had to deal with my somewhat erratic nature on a daily basis. I enjoyed our time together and learned a lot from you during our tea breaks from research and teaching, as well as from our discussions about research and teaching. Among the many others at Tilburg University I want to thank are my friends and colleagues Cansu, Sebastian, Elisabeth, Bas, Sybren, Stefan, Bo, Christos, Chen, Sanja, Renata, Emanuel, Ittai, Jan, Maria, Nick, Ahmadreza, Marleen, Marieke, Krzysztof, Andreas, Manuél, Vatsalya, Peter, Yasir, Tomas, YiLong, Yuxin, Frans, Zhuojiong, Uwe, Victor, Diana, Michal, Clemens and Loes. Our talks and discussions, nights out, bike rides in the

Netherlands and Belgium, or conference trips improved my life in Tilburg tremendously.

Outside my life at University I want to thank Mikael, Joel, Joakim, Joonas, Katarina, Ketty, Felix, Michael, Stefan, Vera, Leah, Leo, Sean, Marit, Velichko and Annelies. You always found a way to take my head off my studies and just enjoy life to its fullest, be it either by going climbing, playing and watching ice hockey, or just sitting together for a beer. Thank you for always being around.

Also, I want to thank my longstanding friend Inna. You were always there for me when I needed it the most, especially after moving to Tilburg when I wanted to quit every other day. You taught me valuable lessons about life and changed my perspective on it.

I would also like to express my deepest thanks to Miriam. The last two and a half years have truely been wonderful. Words cannot express my gratitude towards your encouragement and support, especially during the last stretch of writing this thesis.

Finally, I want to thank my family. My parents Iris and Pius, and my brother Tobias. Without your neverending support, this thesis would not exist. You supported me throughout my life, financed and encouraged my studies from the beginning at the University of Konstanz until the end at Tilburg University. I can, no matter what happens, always count on your support and help. But, more than that, thanks to you I am permanently getting closer to become who I am. Vergelt's Gott!

Mario P. Rothfelder
Tilburg, January 2018

TABLE OF CONTENTS

List of Tables

Table **Page**

List of Figures

CHAPTER 1

PREFACE

This thesis is composed of three essays on time-varying parameters and time series networks where each essay deals with specific aspects thereof. The thesis starts with proposing a 2SLS based test for a threshold in models with endogenous regressors in Chapter 2. Then, Chapter 3 proposes, to my best knowledge, the first estimator for the inverse of the long-run covariance matrix of a linear, potentially heteroskedastic stochastic process. Finally, the thesis concludes with an empirical analysis on the robustness of financial volatility networks with respect to the exclusion of central nodes in Chapter 4.

In Chapter 2, entitled *Testing for a Threshold in Models with Endogenous Regressors* and co-authored with Otilia Boldea, we propose a testing procedure which allows applied researchers to assess whether the data was generated from a single threshold model with endogenous regressors or not. For example, such models can arise when modelling output growth or unemployment rates. To do so, we first outline the class of permissible threshold models and then propose two 2SLS based tests, a sup LR and a sup Wald test. In addition, we derive the asymptotic null distributions of the two tests and show that they depend on the second moment functionals of the data and the chosen functional form of the first stage. However, critical values can easily be simulated via the wild bootstrap. In simulations we found that our tests have empirical size close to the nominal size. This is in sharp contrast to the existing GMM based test of Caner and Hansen (2004) which is severly oversized in small samples. We argue that this occurs for two reasons: First, we theoretically show that estimation via 2SLS can indeed be more efficient than GMM which can result in more accurate empirical sizes. This gain in efficiency is due to the fact that the 2SLS approach utilizes the full infomration contained in the first stage. Secondly, based on our simulation results, we argue heuristically that the wild bootstrap replicates the sample

distributions of the 2SLS tests more accurately than those of the GMM test.

Chapter 3, entitled *Estimating Sparse Long-Run Precision Matrices for Linear Multivariate Time Series*, proposes the first direct estimator for the inverse of the long-run covariance matrix of a potentially heteroskedastic, multivariate linear time series under unknown sparsity constraints. That is, the econometrician does not know which entries of the inverse are equal to zero and which not. Such situations naturally arise, for example, when modelling partial correlation networks based on time series data. The proposed estimator is based on the graphical LASSO of Friedman et al. (2008). That is, the proposed estimator minimizes the ℓ_1-penalized log-likelihood function of $i.i.d.$ multivariate normal data. At first glance this seems counterintuitive since the data is neither $i.i.d.$ nor necessarily normal in a time series setting. However, as I argue one can reinterpret this likelihood function as a special case within the class of Bregman-divergences so that the aforementioned likelihood function measures the distance between any symmetric and positive definite matrix and the true long-run covariance matrix of the underlying process. This interpretation allows me to free the likelihood function from distributional and dependecy assumptions. Since the true long-run covariance matrix is unknown to the econometrician I replace it with a suitable pre-estimator. In particular, I use the HAC estimator with the sharp origin kernel of Phillips et al. (2007). I then show that the resulting adaptive estimator enjoys the oracle property of Zou (2006). That is, the adaptive estimator identifies the zero and non-zero entries with probability tending to one and has the same asymptotic distribution as the oracle estimator. Finally, an extensive Monte Carlo study indicates that the proposed estimator performs well in samples over a wide variety of settings.

Chapter 4, entitled *Robustness of Financial Volatility Networks to the Exclusion of Systemic Nodes*, empirically investigates how robust two commonly applied network measures, the From- and the To-degree, are to the exclusion of central nodes in financial volatility networks. This question is motivated by the current empirical literature which excludes, presumably due to convenience, certain nodes such as Lehman Brothers from their analysis. However, Chapter 3 in Kolaczyk (2017) shows, both theoretically and by simulations, that standard measures of network characteristics are biased in unknown directions when nodes of the network are excluded. Therefore, this chapter aims to assess to what extend the exclusion of Lehman Brothers, decidedly an important node in the U.S. financial system, affects the aforementioned network measures and, thereby, possibly distorts current empirical results and conclusions. To do so, I make use of the most commonly applied network in the literature, the long-run variance decomposition network of Diebold and Yilmaz (2014). I estimate this network based on a VAR(1)-representation of the data, once when Lehman Brothers is excluded and once when Lehman Brothers is included since this allows me to gauge the effects Lehman Brothers' stock has on the From- and To-degree network measures. I find that the To-degree is heavily affected by the exclusion of Lehman

Brothers whereas the From-degree seems to be only minorly affected. These results hold on a firm-specific and aggregated sector level for a sparse and non-sparse VAR-representation of the data.

CHAPTER 2

TESTING FOR A THRESHOLD IN MODELS WITH ENDOGENOUS REGRESSORS

This chapter is based on the identically entitled working paper which is co-authored with Otilia Boldea

Using 2SLS estimation, we propose two tests for a threshold in models with endogenous regressors: a sup LR test and a sup Wald test. Here, the 2SLS estimation is not conventional because it uses additional information about the first-stage being linear or not. Because of this additional information, our tests can be more accurate than the threshold test in Caner and Hansen (2004) which is based on conventional GMM estimation.

We derive the asymptotic distributions of the two tests for a linear and for a threshold first stage. In both cases, the distributions are non-pivotal, and we propose obtaining critical values via a fixed regressor wild bootstrap. Our simulations show that in small samples, the GMM test of Caner and Hansen (2004) can be severely oversized under heteroskedasticity, while both the 2SLS tests we propose are much closer to their nominal size. Therefore, we recommend using both our tests in small samples, to avoid detecting a threshold when there is none.

2.1 Introduction

Threshold models are widely used in economics to model unemployment, output, growth, bank profits, asset prices, exchange rates, and interest rates. See Hansen (2011) for a survey of economic applications.

Pioneered by Howell Tong - see e.g. Tong (1990), threshold models with exogenous regressors have been widely studied and their asymptotic theory is well known.[1] Even though exogeneity is violated in many economic applications, papers on threshold regression with *endogenous regressors* remain relatively scarce. They were pioneered by Caner and Hansen (2004), who show that when regressors are endogenous but the threshold variable is exogenous, the threshold parameter can be estimated by minimizing a two stage least squares (2SLS) criterion over values of the threshold variable encountered in the sample.

In general, the applied researcher needs to decide whether there is a threshold to begin with. This can be done via testing for an unknown threshold. For example, the government spending multiplier is often conjectured to be larger in regimes where the nominal interest rate is close to the zero lower bound - see Eggertsson (2010) and Christiano et al. (2011).[2] This conjecture can be validated by testing whether there is a threshold driven by low interest rates. Another example is testing whether growth slows down when the debt to GDP ratio is high - see Reinhart and Rogoff (2010) (tests for this conjecture albeit using exogenous regressors can be found in Lee et al. (2014) and Hansen (2016) a.o.). Many more examples can be found in Hansen (2011).

In this chapter, we develop 2SLS tests for no threshold against the alternative of one unknown threshold for models with endogenous regressors. Caner and Hansen (2004) already proposed a GMM sup Wald test for the same hypothesis. Here, we show that this test is severely oversized in small, heteroskedastic samples. We propose instead two 2SLS tests (a 2SLS sup LR test and a 2SLS sup Wald test), which we show have superior size properties in finite samples. The superior size stems from how the 2SLS estimators are constructed. They are not conventional, because they use additional information about the first stage, while the conventional GMM estimators in Caner and Hansen (2004) do not use any information about the first stage. With this additional information, we show that the 2SLS estimators can be more accurate than the conventional GMM estimators, and that they lead to better sized tests in finite samples.[3]

The additional information we use is whether there is a threshold in the first stage. We consider two cases: the first stage is a linear model and the first stage is a threshold model.[4] We

[1] See a.o. Hansen (1996, 1999, 2000) and Gonzalo and Wolf (2005) for inference, Gonzalo and Pitarakis (2002) for multiple threshold regression and model selection, Caner and Hansen (2001) and Gonzalo and Pitarakis (2006) for threshold regression with unit roots, Seo and Linton (2007) for smoothed estimators of threshold models, Lee et al. (2011) for testing for thresholds, and Hansen (2016) for threshold regressions with a kink.

[2] This can happen because when the monetary policy is less effective, fiscal stimulus can quickly lower real interest rates by raising inflation, resulting in potentially large multiplier effects.

[3] These unconventional 2SLS estimators were already proposed in Caner and Hansen (2004), but not for constructing tests for a threshold.

[4] Caner and Hansen (2004) consider the same cases for estimating the threshold parameter, but not for testing for

compute the 2SLS tests for each case separately, and show that their null asymptotic distributions depend on the data and on the case considered. Nevertheless, critical values are straightforward to compute via the wild bootstrap, so these tests are easily implemented in practice. To our knowledge, this is the first paper to propose and analyze 2SLS tests for a threshold.

We study the properties of both tests via simulation. We generate critical values via a fixed regressor wild bootstrap that we describe in this paper. We find that the 2SLS sup LR and the 2SLS sup Wald test are either correctly sized or slightly undersized. In contrast, the GMM sup Wald test is correctly sized under homoskedasticity, but under heteroskedasticity, it is severely oversized.[5] This holds for both linear and threshold first stages. As the sample size grows large, both our tests approach their nominal sizes, and the GMM test does too, albeit slower than our tests. Since we find no systematic difference between the two 2SLS tests, we conclude that both are valuable alternative diagnostics to the GMM test for a threshold, especially under heteroskedasticity.

The chapter is closely related to two papers in the break-point literature - Hall et al. (2012) and Boldea et al. (2017). Both papers study the 2SLS sup LR and 2SLS sup Wald tests for a break, the first one for a linear first stage, the second one for a first stage with a break. The asymptotic distributions for the break-point tests are pivotal in the first paper and depend on the break in the first stage in the second paper. In contrast, we find that the asymptotic distributions of the threshold tests are non-pivotal in both cases, a linear or a threshold first stage. Moreover, they are very different than the break-point distributions, and we show that they only coincide in unrealistic threshold models.

The chapter is also related to Magnusson and Mavroeidis (2014), who use information about break-points in the first stage (and in general break-points in the derivative of the moment conditions) to improve efficiency of tests for moment conditions. It is also related to Antoine and Boldea (2017) and Antoine and Boldea (2015): the first uses breaks in the Hessian of the GMM minimand and the second uses full sample FS information. Both papers focus on more efficient estimation, while we focus on improved testing.

It should be noted that we allow for endogenous regressors, but not for endogenous threshold variables. For the latter, see Kourtellos et al. (2015). Also, to account for regressor endogeneity, we make use of instruments for constructing parametric test statistics for thresholds. As a result, our tests have nontrivial local power for $O(T^{-1/2})$ threshold shifts. This is in contrast with Yu and Phillips (2014), who does not use instruments, but rather local shifts around the threshold to construct a nonparametric threshold test. As a result, his test covers more general models, at the cost of losing power in $O(T^{-1/2})$ neighborhoods.

a threshold. One can distinguish between the two cases by testing for a threshold in the first stage, using currently available tests such as the OLS sup Wald test in Hansen (1996).

[5]Note that unlike the Wald test for classical hypotheses, the (heteroskedasticity-robust) sup Wald test for the null hypothesis for an unknown threshold does not have a pivotal null distribution. That means that correcting for heteroskedasticity (and therefore using Wald tests instead of LR tests) does not necessarily result in better size properties for the sup Wald test compared to the sup LR test; this is indeed what we find in the simulations.

This chapter is organized as follows. Section 2.2 introduces the threshold model. Section 2.3 defines the 2SLS and GMM estimators, and theoretically and numerically motivates the use of 2SLS estimators. Section 2.4 defines the new 2SLS test statistics and derives their asymptotic distributions. Section 2.5 describes the existing GMM test of Caner and Hansen (2004). Section 2.6 describes the fixed regressor wild bootstrap, and illustrates the small sample properties of all tests via simulations. Section 2.7 concludes. All the proofs are relegated to the Appendix, together with additional notation.

2.2 Threshold Model

Our framework is a linear model with a possible threshold at γ^0:

$$\begin{aligned} y_t &= \left(z_t^\top \theta_{1z}^0 + x_{1t}^\top \theta_{1x}^0\right) \mathbb{1}_{\{q_t \le \gamma^0\}} + \left(z_t^\top \theta_{2z}^0 + x_{1t}^\top \theta_{2x}^0\right) \mathbb{1}_{\{q_t > \gamma^0\}} + \epsilon_t \\ &= w_t^\top \theta_1^0 \mathbb{1}_{\{q_t \le \gamma^0\}} + w_t^\top \theta_2^0 \mathbb{1}_{\{q_t > \gamma^0\}} + \epsilon_t. \end{aligned}$$

Here, y_t is the dependent variable, z_t is a $p_1 \times 1$-vector of endogenous variables and x_{1t} a $p_2 \times 1$-vector of exogenous variables containing the intercept, and $w_t = (z_t^\top, x_{1t}^\top)^\top$. We set $p_1 + p_2 = p$. Also, q_t is the exogenous threshold variable (which can be a function of the exogenous regressors) and $\mathbb{1}_{\{\mathscr{A}\}}$ denotes the indicator function on the set $\mathscr{A}$. Furthermore, for $i = 1,2$, θ_{iz}^0 are $p_1 \times 1$-vectors of slope parameters associated with z_t, θ_{ix}^0 are $p_2 \times 1$-vectors of the slope parameters associated with x_{1t} and $\gamma^0 \in \Gamma^0 = [\gamma_{min}, \gamma_{max}]$, its compact support.[6] The second equation is just a more compact way of writing the first, with $w_t = (z_t^\top, x_{1t}^\top)^\top$ being the augmented regressors, and $\theta_i^0 = (\theta_{iz}^{0\top}, \theta_{ix}^{0\top})^\top$ being $p \times 1$-vectors of the slope parameters, for $i = 1,2$.

We assume that z_t is endogenous ($\mathbb{E}[\epsilon_t] = 0$; $\mathbb{E}[z_t \epsilon_t] \neq 0$) and strong instruments x_t are available; these instruments include x_{1t}, the exogenous regressors.

As in Caner and Hansen (2004), we consider two different specifications for the first stage FS: a linear first stage (LFS), given by

$$z_t = \Pi^{0\top} x_t + u_t,$$

and a threshold first stage (TFS) given by

$$z_t = \Pi_1^{0\top} x_t \mathbb{1}_{\{q_t \le \rho^0\}} + \Pi_2^{0\top} x_t \mathbb{1}_{\{q_t > \rho^0\}} + u_t.$$

In both specifications for the FS, $x_t = (x_{1t}^\top, x_{2t}^\top)^\top$ is a $q \times 1$-vector with $q \ge p$, $q = p_2 + q_1$; Π^0, Π_1^0 and Π_2^0 are $q \times p_1$-matrices of the FS slope parameters; $\rho^0 \in \Gamma^0$ is the FS threshold parameter, possibly different than γ^0, with the same support Γ^0.

As common in the threshold literature, we assume that ϵ_t and u_t are martingale differences, i.e. $\mathbb{E}[\epsilon_t | \mathfrak{F}_t] = 0$ and $\mathbb{E}[u_t | \mathfrak{F}_t] = \mathbf{0}$, $\mathfrak{F}_t = \sigma\{q_{t-s}, x_{t-s}, u_{t-s-1}, \epsilon_{t-s-1} | s \ge 0\}$, and $(x_t^\top, z_t^\top)^\top$ is measurable

[6] We can allow for $\Gamma^0 = \mathbb{R}$. However, the end-points of the support of q_t, even when infinite, are relevant for simulating asymptotic $p-$values. Without further information, the only end-points we observe are those in the sample: the minimum and maximum value of q_t, which we call $\gamma_{min}, \gamma_{max}$; therefore, we fix $\Gamma^0 = [\gamma_{min}, \gamma_{max}]$.

with respect to $\mathfrak{F}_t$. This assumption implies that the threshold variable q_t is exogenous, and so are the instruments x_t.

Next, we write the equations above in matrix form. To do so, stack all observations in the following T-row matrices:

$$X_1^\rho = \left(x_t^\top \mathbb{1}_{\{q_t \le \rho\}}\right)_{t=1,\dots,T} \quad X_2^\rho = \left(x_t^\top \mathbb{1}_{\{q_t > \rho\}}\right)_{t=1,\dots,T}$$
$$W_1^\gamma = \left(w_t^\top \mathbb{1}_{\{q_t \le \gamma\}}\right)_{t=1,\dots,T} \quad W_2^\gamma = \left(w_t^\top \mathbb{1}_{\{q_t > \gamma\}}\right)_{t=1,\dots,T}.$$

Let Y, X, Z, W, ϵ and u be the matrices stacking observations $t = 1,\dots,T$. Then the LFS is:

$$Z = X\Pi^0 + u \tag{2.1}$$

and the TFS is:

$$Z = X_1^{\rho^0}\Pi_1^0 + X_2^{\rho^0}\Pi_2^0 + u. \tag{2.2}$$

The equation of interest - which can arise from a structural model and for lack of better terminology is called the equation of interest (EI) - is, for a threshold parameter γ^0:

$$Y = W_1^{\gamma^0}\theta_1^0 + W_2^{\gamma^0}\theta_2^0 + \epsilon. \tag{2.3}$$

If there is no EI threshold, $\theta_1^0 = \theta_2^0 = \theta^0$, and the EI is $Y = W\theta^0 + \epsilon$.

Note that we allow for the case of a threshold in the first stage without any threshold in the equation of interest. For example, if the equation of interest is a structural model where inflation depends endogenously on output, there can be different output regimes that do not affect the structural parameters of the inflation model over extended periods, as shown empirically in Antoine and Boldea (2017). Similarly, we allow for the equation of interest to have a threshold when the first stage has none. For example, if the equation of interest is a monetary policy rule where interest rates are targeting the endogenous inflation, we may have regime shifts in the policy rule without the first stage equation for inflation being affected - see Antoine and Boldea (2015). Even if there is a threshold in both the equation of interest and its first stage, the values of the threshold need not coincide, for example, because the policy modelled in the first stage reacts to deteriorating business conditions differently than the real economy modelled in the second stage or equation of interest.

2.3 2SLS versus GMM estimation

In this section, we motivate the use of 2SLS estimation for constructing test statistics. We are interested in testing for a EI threshold, the null hypothesis being $\mathbb{H}_0 : \theta_1^0 = \theta_2^0$ in (2.3). Because γ^0 is usually unknown and it is a nuisance parameter under the null hypothesis, a common practice is to calculate a series of test statistics, each for a given $\gamma \in \Gamma$ (where $\Gamma \subset \Gamma^0$), and then to take the supremum over these quantities to obtain a single test statistic for the null of no threshold

against the alternative of one threshold. For example, Hansen (1996) and Caner and Hansen (2004) construct such tests.

In the presence of endogenous regressor, to test for $\mathbb{H}_0$, Caner and Hansen (2004) defines two-step GMM estimators of $\theta_i^0, (i = 1,2)$ for each γ. These are conventional in the sense that by construction, they ignore any information about the FS. Specifically, for each $\gamma \in \Gamma$, where Γ is a closed interval in the support Γ^0, bounded away from the end-points of this support, and $i = 1,2$:

$$\hat{\theta}_{i,GMM}^{\gamma} = \left(W_i^{\gamma\top} X_i^{\gamma} \hat{H}_{i,GMM}^{\epsilon^{-1}}(\gamma) X_i^{\gamma\top} W_i^{\gamma}\right)^{-1} \left(W_i^{\gamma\top} X_i^{\gamma} \hat{H}_{i,GMM}^{\epsilon}(\gamma) X_i^{\gamma\top} Y\right),$$

with estimated long-run variances:

$$\hat{H}_{1,GMM}^{\epsilon}(\gamma) = T^{-1} \sum_{t=1}^{T} \hat{\epsilon}_{t,GMM}^2 x_t x_t^{\top} \mathbb{1}_{\{q_t \leq \gamma\}}, \hat{H}_{2,GMM}^{\epsilon}(\gamma) = T^{-1} \sum_{t=1}^{T} \hat{\epsilon}_{t,GMM}^2 x_t x_t^{\top} \mathbb{1}_{\{q_t > \gamma\}},$$

where $\hat{\epsilon}_{t,GMM}$ is the t^{th} element of the $T \times 1$ vector $\hat{\epsilon}_{GMM} = y - W_1^{\gamma} \tilde{\theta}_{1,GMM}(\gamma) - W_2^{\gamma} \tilde{\theta}_{2,GMM}(\gamma)$, and $\tilde{\theta}_{i,GMM}(\gamma)$ are some preliminary first step GMM estimators of (2.3) for a given γ and $i = 1,2$.[7]

If instead, we estimate (2.3) by 2SLS, we have no choice but to take into account the nature of the FS - linear model or threshold model - otherwise the resulting estimator of θ_i^0 may be inconsistent. These two cases - linear or threshold FS - have also been considered in Caner and Hansen (2004) for 2SLS slope estimators, but with the purpose of defining a consistent estimator the threshold parameter γ^0.

For a linear FS (LFS), let:

$$\hat{Z} = X\hat{\Pi}, \ \hat{W} = \left(\hat{Z}, X_1\right), \tag{2.4}$$

with $X_1 = (x_{1t}^{\top})_{t=1,...,T}$.

For a threshold FS (TFS), first estimate the threshold parameter ρ as in Caner and Hansen (2004):

$$\hat{\rho} = \underset{\rho \in \Gamma}{\operatorname{argmin}} \det\left(\hat{u}(\rho)^{\top} \hat{u}(\rho)\right), \tag{2.5}$$

where $\hat{u}(\rho) = Z - X_1^{\rho} \hat{\Pi}_1(\rho) - X_2^{\rho} \hat{\Pi}_2(\rho)$ and $\hat{\Pi}_1(\rho), \hat{\Pi}_2(\rho)$ are the OLS estimators of Π_1^0, Π_2^0 in (2.2) for a given ρ:

$$\hat{\Pi}_1(\rho) = \left(X_1^{\rho\top} X_1^{\rho}\right)^{-1} X_1^{\rho\top} Z \tag{2.6}$$

$$\hat{\Pi}_2(\rho) = \left(X_2^{\rho\top} X_2^{\rho}\right)^{-1} X_2^{\rho\top} Z. \tag{2.7}$$

With $\hat{\rho}$, the TFS slope parameter estimates are $\hat{\Pi}_1 = \hat{\Pi}_1(\hat{\rho})$, $\hat{\Pi}_2 = \hat{\Pi}_2(\hat{\rho})$.

Then:

$$\hat{Z} = \hat{\Pi}_1 X_1^{\hat{\rho}} + \hat{\Pi}_2 X_2^{\hat{\rho}}. \tag{2.8}$$

[7]Note that because W are already partitioned according to $\mathbb{1}_{\{q_t \leq \gamma\}}$, we have $W_i^{\gamma\top} Y = W_i^{\gamma\top} Y_i$.

The second-stage of the 2SLS is standard. Construct $\hat{W} = \left(\hat{Z}, X_1\right)$, with $\hat{Z}$ defined in (2.4) for a LFS and (2.8) for a TFS, and the 2SLS estimators of θ_1^0, θ_2^0 for a given $\gamma \in \Gamma$ are for $i = 1,2$.

$$\hat{\theta}_1^\gamma = \left(\hat{W}_1^{\gamma\top}\hat{W}_1^\gamma\right)^{-1}\left(\hat{W}_1^{\gamma\top}Y\right) \tag{2.9}$$

$$\hat{\theta}_2^\gamma = \left(\hat{W}_2^{\gamma\top}\hat{W}_2^\gamma\right)^{-1}\left(\hat{W}_2^{\gamma\top}Y\right). \tag{2.10}$$

Next, we provide two reasons why we advocate the use of 2SLS over GMM when one is interested in deciding whether a threshold is present in the EI or not. One is theoretical and provides an argument that the 2SLS estimators for θ_i^γ, $i = 1,2$, can be more efficient than GMM under $\mathbb{H}_0$ and the second is a heuristic argument based on results from our Monte Carlo simulations where we find that the bootstrapped distributions of the 2SLS test statistics are a better fit to the empirical distributions than in case of GMM.

Efficiency Both the 2SLS and the GMM estimators defined here are consistent under standard assumptions, as shown in Caner and Hansen (2004). But the GMM estimators ignore potentially valid information about the FS. As a result, the GMM estimators can be less efficient than the 2SLS estimators which, in turn, can distort the empirical sizes of a GMM-based threshold test. This result is formalized below.

Theorem 2.1 (2SLS versus GMM)**.**
Assume the EI is (2.3) *with the TFS* (2.2)*, one endogenous regressor, one instrument and no exogenous regressors ($p = q = p_1 = 1$), and impose $\mathbb{H}_0$: $\theta_z^0 = \theta_{1z}^0 = \theta_{2z}^0$. Let ρ^0 be known and let Assumptions 2.1–2.4 of Section 2.4.2 hold, with $\sigma_\epsilon^2 = \mathrm{Var}(\epsilon_t)$ and $\sigma^2 = \mathrm{Var}(\epsilon_t + u_t\theta_z^0)$. Then, for a given γ,*
(i) For both $i = 1,2$,

$$\sqrt{T}(\hat{\theta}_i^\gamma - \theta^0) \xrightarrow{d} \mathcal{N}(0, V_{A,i}^*(\gamma)) \text{ and } \sqrt{T}(\hat{\theta}_{i,GMM}^\gamma - \theta^0) \xrightarrow{d} \mathcal{N}(0, V_{i,GMM}^*(\gamma)),$$

where $V_{A,i}^*(\gamma)$ and $V_{i,GMM}^*(\gamma)$ are defined in Lemma 2.B.9 of the Appendix.
(ii) If $\sigma^2 \le \sigma_\epsilon^2$, then $\left\{V_{i,GMM}^*(\gamma) \ge V_{A,i}^*(\gamma) \text{ for both } i = 1,2 \text{ simultaneously}\right\}$.
(iii) If the FS is in fact linear, that is, if $\Pi_1^0 = \Pi_2^0$, then:

$$\sigma^2 \le \sigma_\epsilon^2 \iff \left\{V_{i,GMM}^*(\gamma) \ge V_{A,i}^*(\gamma) \text{ for both } i = 1,2 \text{ simultaneously}\right\}$$

(iv) $V_{i,GMM}^*(\rho^0) = V_{A,i}^*(\rho^0)$.

Note that Theorem 2.1 is derived under conditional homoskedasticity (imposed in Assumption 2.2) and under independence of q_t and x_t (imposed in Assumption 2.3).[8]

The intuition for the results in Theorem 2.1 is as follows. If *the sample $\{t : q_t \le \gamma\}$ is used for both the FS and the EI* to compute 2SLS estimators, and the *same sample* is used for

[8] In more general cases, it is much harder to obtain a similar result analytically.

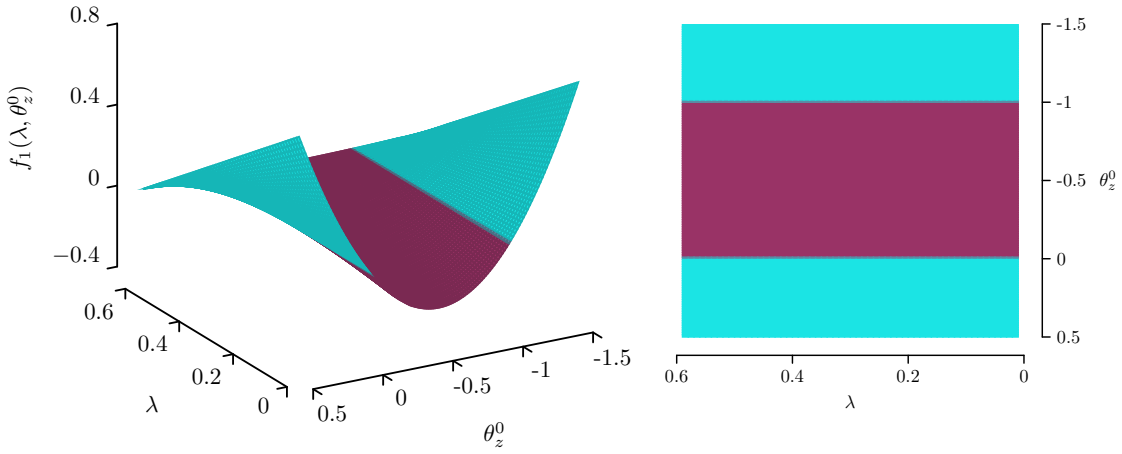
GMM estimators, then both these estimators are conventional. Therefore, the two-step GMM is asymptotically more efficient than the 2SLS, and asymptotically equivalent in the just-identified case. This is shown in Theorem 2.1(iv) where we set $\gamma = \rho^0$. However, when $\gamma \neq \rho^0$ the 2SLS estimators are not conventional. For example, if $\gamma \leq \rho^0$, in computing the 2SLS estimator over the sample $\{t : q_t \leq \gamma\}$, we *use information from the FS over a larger sample* $\{t : q_t \leq \rho^0\}$. Theorem 2.1 (ii) shows that this additional information leads to more efficient estimators if the 2SLS errors $(\epsilon_t + u_t\theta_z^0)$ have smaller variance than the GMM errors ϵ_t. This efficiency result also holds if instead the FS is linear, as shown in Theorem 2.1(iii).

Theorem 2.1 is not just a theoretical result, as shown in the example below.

Example 2.1. *Suppose that* $\Pi_1^0 = 1$, $\Pi_2^0 = 1.25$, $\rho^0 = 0.25$. *Let* $q_t \overset{iid}{\sim} \mathcal{N}(0,1)$, $x_t \overset{iid}{\sim} \mathcal{N}(0,1)$ *and* $\begin{bmatrix} \epsilon_t \\ u_t \end{bmatrix} \overset{iid}{\sim} \mathcal{N}\left(0, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$. *Let* $f_i(\lambda, \theta_z^0) = V_{A,i}^*(\gamma) - V_{GMM,i}^*(\gamma)$, *and* $\gamma \leq \rho^0$ *(if* $\gamma > \rho^0$, *the first plot becomes the second and viceversa).*

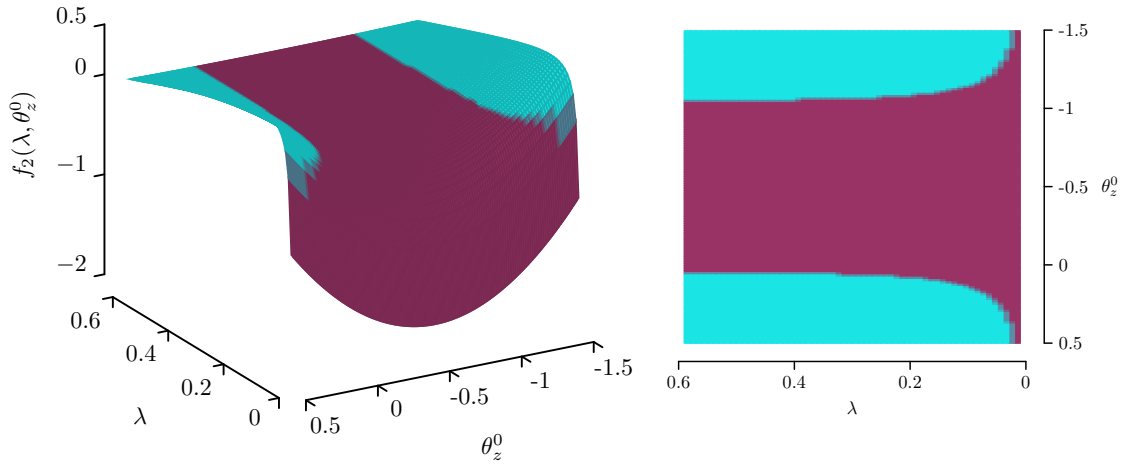
Note that in this case, $\sigma^2 - \sigma_\epsilon^2 = (\theta_z^0)(1+\theta_z^0)$. From Theorem 2.1, if $\theta_z^0(1+\theta_z^0) < 0$, $f_i(\lambda, \theta_z^0) < 0$ and both 2SLS estimators are more efficient.[9] From Example 1, $\mu^0 \equiv \mathbb{E}\mathbb{1}_{\{q_t \leq \rho^0\}} = 0.5981$. In Figures 2.1 and 2.2 we plot $f_1(\lambda, \theta_z^0)$ and $f_2(\lambda, \theta_z^0)$ as functions of $\theta_z^0 \in [-1.5, 0.5]$ and $\lambda = P(q_t \leq \gamma) \in (0, \mu^0]$. The purple areas indicate parameter configurations where 2SLS is more efficient than GMM, and these are sizable areas of the parameter space.

Figure 2.1: Plot and Contour Plot of $f_1(\cdot)$



[9] As shown in the proof of Theorem 2.1, when $\sigma^2 > \sigma_\epsilon^2$, $\hat{\theta}_1^\gamma$ is less efficient than $\hat{\theta}_{1,GMM}^\gamma$, but $\hat{\theta}_2^\gamma$ can still be more efficient than $\hat{\theta}_{2,GMM}^\gamma$ depending on the DGP.

Figure 2.2: Plot and Contour Plot of $f_2(\cdot)$



Bootstrap Accuracy The second argument why our 2SLS tests should be preferred over the GMM test is heuristic in nature and motivated by our findings from the simulation study in Section 2.6. For the sake of brevity, we consider the LFS case of Section 2.6 for three cases: homoskedasticity that is known to the researcher, homoskedasticity that is unknown to the researcher, and heteroskedasticity.[10] Figure 2.3 plots the empirical and bootstrapped distributions of the 2SLS and GMM test statistics for these three cases.

In the first case, we know that the errors are homoskedastic and use this information both for the bootstrap and for the construction of the test statistics, the bootstrapped distributions closely matches the empirical distributions, so all three tests are equally well sized.

In the second case, we do not know that the errors are homoskedastic and, therefore, we use the wild bootstrap and heteroskedasticity-robust test statistics. In this case, the bootstrapped distribution of the GMM test no longer closely matches the empirical distribution. This is especially pertinent in the right tail of the distributions, which is associated with the critical values of the test statistic. In contrast, the bootstrapped distributions continue to closely match the empirical distributions for the 2SLS tests. Therefore, the 2SLS tests provide the researcher with more accurate decisions about the presence of a threshold in the EI than the existing GMM test. Moreover, these results are robust to using different estimators for the heteroskedasticity robust covariances (known as HCCME0–3) and to using different forms of the wild bootstrap.[11]

In the third case, we have heteroskdasticity; this did not change the results of the second

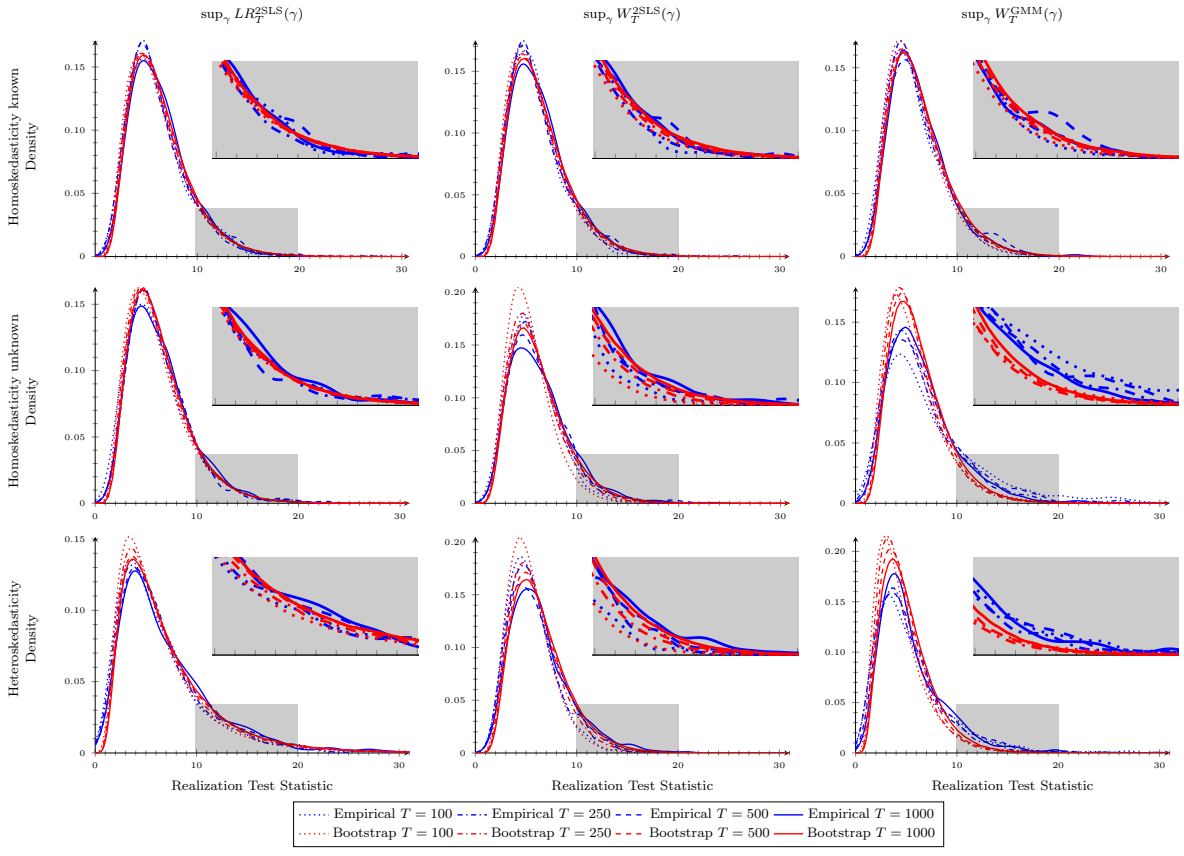
[10] As we will discuss in Section 2.6, when we know that the errors are homoskedastic we replace the wild bootstrap by the $i.i.d.$ bootstrap where we re-sample the error terms from a (multivariate) normal distribution with mean zero and variance given by the sample variance of the residuals. Moreover, we adjust all test statistics so that they incorporate information about homoskedasticity. That is, we replace quantities of the form $\mathbb{E}[x_t x_t^\top \epsilon_t^2]$ by $\sigma_\epsilon^2 \mathbb{E}[x_t x_t^\top]$, etc. If we do not know that the errors are homoskedastic then we use the wild bootstrap and the heteroskedasticity-robust test statistics.

[11] These results are available from the authors upon request.

case, even when varying the skedastic function and the degree of heteroskedasticity. Finally, the same applies when we consider the TFS case.[12]

Tables 2.1 - 2.4 reinforce the results discussed above for both a LFS and a TFS. They show that the GMM test is severely oversized in small samples; at a nominal size of 5%, the empirical sizes reach up to 15% for 100 observations; they decrease as the sample size increases, but they are still around 6 – 10% for 1000 observations. Since many applications of threshold tests are macroeconomic applications, where a representative sample is around 500 observations, the size distortions of the GMM test are worrisome, as they will often lead to favor a threshold model when the true model is linear. The same tables show that the 2SLS tests are either correctly sized or slightly undersized, but not oversized. This motivates us to consider the 2SLS tests as complementary threshold diagnostics.

Figure 2.3: Empirical and bootstrapped distributions of the 2SLS and GMM test statistics.



[12]These results are available from the authors upon request.

2.4 2SLS Tests

2.4.1 Test Statistics

For a LFS, the first test statistic we propose is a sup LR test in the spirit of Davies (1977):

$$\sup_{\gamma\in\Gamma} LR^{2SLS}_{T,LFS}(\gamma) = \sup_{\gamma\in\Gamma} \frac{SSR_0 - SSR_1(\gamma)}{SSR_1(\gamma)/(T-2p)}, \tag{2.11}$$

where SSR_0 and $SSR_1(\gamma)$ are the 2SLS sum of squared residuals under the null and the alternative hypotheses:

$$\begin{aligned} SSR_0 &= (Y - \hat{W}\hat{\theta})^\top (Y - \hat{W}\hat{\theta}), \\ SSR_1(\gamma) &= (Y_1^\gamma - \hat{W}_1^\gamma \hat{\theta}_1^\gamma)^\top (Y_1^\gamma - \hat{W}_1^\gamma \hat{\theta}_1^\gamma) + (Y_2^\gamma - \hat{W}_2^\gamma \hat{\theta}_2^\gamma)^\top (Y_2^\gamma - \hat{W}_2^\gamma \hat{\theta}_2^\gamma), \end{aligned}$$

and where $\hat{\theta} = (\hat{W}^\top \hat{W})^{-1} \hat{W}^\top Y$ is the full-sample 2SLS estimator, and $\hat{W}, \hat{\theta}_1^\gamma, \hat{\theta}_2^\gamma$ are defined in Section 2.3 for a LFS.

A scaled version of this test is known as the sup F test in the break-point literature - see Bai and Perron (1998) for OLS and Hall et al. (2012) for 2SLS.

We also propose the sup Wald test:

$$\sup_{\gamma\in\Gamma} W^{2SLS}_{T,LFS}(\gamma) = \sup_{\gamma\in\Gamma} T \left[\hat{\theta}_1^\gamma - \hat{\theta}_2^\gamma\right]^\top \hat{V}^{-1}(\gamma) \left[\hat{\theta}_1^\gamma - \hat{\theta}_2^\gamma\right], \tag{2.12}$$

where $\hat{V}(\gamma)$ is defined in Definition 2.2 of the Appendix, and unlike the 2SLS sup Wald test in Hall et al. (2012), it takes into account that the 2SLS estimators $\hat{\theta}_1^\gamma$ and $\hat{\theta}_2^\gamma$ are correlated through a full-sample first-stage.

For a TFS, the test statistics are calculated exactly as above, but taking into account the TFS when computing the first stage of the 2SLS estimation, as in (2.8). Therefore, $\sup_{\gamma\in\Gamma} W^{2SLS}_{T,TFS}(\gamma)$ is computed with $\hat{V}_A(\gamma)$ instead of $\hat{V}(\gamma)$, and $\hat{V}_A(\gamma)$ is defined in Definition 2.3 of the Appendix.

2.4.2 Assumptions

Define

$$M_1(\gamma) = \mathbb{E}[x_t x_t^\top \mathbb{1}_{\{q_t \le \gamma\}}], \;\; M = M(\gamma_{\max}) = \mathbb{E}[x_t x_t^\top], \text{and} \;\; M_2(\gamma) = M - M_1(\gamma)$$

as the second moment functionals of the instruments x_t, where $\gamma \in \Gamma$. We impose similar but slightly stronger assumptions than in Caner and Hansen (2004) below, mainly for clarity of our proofs.

Assumption 2.1.

1. *Let $v_t = (\epsilon_t, u_t^\top)^\top$ denote the compound error term. Then*

$$\mathbb{E}[v_t | \mathfrak{F}_t] = 0$$

with $\mathfrak{F}_t = \sigma\{x_{t-s}, v_{t-s-1}, q_{t-s} | s \ge 0\}$.

2. *The series* $(\epsilon_t, u_t^\top, x_t^\top, z_t^\top, q_t)^\top$ *is strictly stationary and ergodic with* ρ*-mixing coefficient* $\rho(m) = \mathcal{O}(m^{-A})$ *for some* $A > \frac{a}{a-1}$ *and* $1 < a \leq 2$*. Also, for some* $b > a$*,*

$$\sup_t \mathbb{E}\|x_t\|_2^{4b} < \infty, \ \sup_t \mathbb{E}\|v_t\|_2^{4b} < \infty,$$

with $\|\cdot\|_2$ *being the Euclidean norm, and* $\inf_{\gamma \in \Gamma} \det M_1(\gamma) > 0$*.*

3. *The density of* v_t *is absolutely continuous, bounded and positive everywhere.*

4. *The threshold variable* q_t *has a continuous pdf* $f(q_t)$ *with* $\sup_t |f(q_t)| < \infty$*.*

5. *The variance of the compound error term* v_t *is given by*

$$\mathbb{E}[v_t v_t^\top] = \Sigma = \begin{pmatrix} \sigma_\epsilon^2 & \Sigma_{\epsilon,u}^\top \\ \Sigma_{\epsilon,u} & \Sigma_u \end{pmatrix},$$

which is positive definite.

6. *Assume* Π^0 *(LFS) or* Π_1^0, Π_2^0 *(TFS) are full rank.*

2.1.1 is needed for threshold models, and it excludes autocorrelation in the errors. However, lagged regressors can enter both the EI and the FS. 2.1.2 is standard for time series and is trivially satisfied for many cross-section models (note that even though we use the time series notation with index t, our results equally apply to cross section models). However, it precludes nonstationary processes. 2.1.3 is needed in the TFS case in order to make asymptotic statements about the FS parameters in the spirit of Chan (1993). 2.1.4 requires the support of q_t to be continuous; if it is discrete, the search over Γ is much easier to perform. 2.1.5 allows conditional heteroskedastic errors and finally, 2.1.6 says that x_t is a strong instrument.

Assumption 2.2.

$$\mathbb{E}[v_t v_t^\top | \mathfrak{F}_{t-1}] = \Sigma = \begin{pmatrix} \Sigma_\epsilon & \Sigma_{\epsilon,u}^\top \\ \Sigma_{\epsilon,u} & \Sigma_u \end{pmatrix}.$$

Assumption 2.2 is a conditional homoskedasticity assumption, which we only use for special case derivations.

Assumption 2.3. *The threshold variable* q_t *and the vector of exogenous variables* x_t *are independent. i.e.*

$$q_t \perp x_t \ \ \forall t = 1, 2, ..., T.$$

Assumption 2.3 is also quite strong and is only used to relate the results in this paper to those on break-point tests, not for the main results of the paper. It doesn't allow the threshold variable q_t to be one of the instrumental variables or exogenous regressors x_t, and is quite restrictive. However, it mimics break-point models, where the threshold is time, or more exactly, a fraction of the sample size, t/T.

Assumption 2.4 (Identifiability)**.** *If we have a TFS as in* (2.2)*,* $\Pi_1^0 \neq \Pi_2^0$.

Assumption 2.4 states that if there is a TFS, the threshold effect is large. It is imposed for simplicity.

2.4.3 Asymptotic distributions with a LFS

To write the asymptotic distributions, define the "ratios"

$$R_i(\gamma) = M_i(\gamma)M^{-1}, i = 1,2.$$

Also, let

$$\mathscr{GP}_{\text{mat},1}(\gamma) \quad \text{and} \quad \mathscr{GP}_{\text{mat}}$$

as $q \times (p_1+1)$-matrices where all columns are $q \times 1$ zero mean Gaussian processes, and the covariance kernels of $\mathscr{GP}_1(\gamma) = \text{vec}(\mathscr{GP}_{\text{mat},1}(\gamma))$ and $\mathscr{GP} = \text{vec}(\mathscr{GP}_{\text{mat}})$ are given by $\mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t \leq \gamma\}}]$ and $\mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)]$. Let $\mathscr{GP}_{\text{mat}} = \mathscr{GP}_{\text{mat},1}(\gamma_{\max})$.
Also, let

$$A^0 = [\Pi^0, S^\top]^\top$$

be the augmented matrix of the FS slope parameters, where $S = [I_{p_2}, \mathbf{0}_{p_2 \times q_1}]$, I_{p_2} is the $p_2 \times p_2$ identity matrix and $\mathbf{0}_{p_2 \times q_1}$ a $p_2 \times q_1$ null matrix ($p_2 + q_1 = q$). Hence, $x_{1t} = Sx_t$ and $w_t = A^0 x_t + \bar{u}_t$, where $\bar{u}_t = (u_t^\top, \mathbf{0}_{1 \times q_1})^\top$. Define the matrices

$$C_1(\gamma) = A^0 M_1(\gamma) A^{0\top}, \quad C = C_1(\gamma_{\max}) = A^0 M A^{0\top}, \quad \text{and} \quad C_2(\gamma) = C - C_1(\gamma)$$

and the Gaussian process:

$$\mathscr{B}_1(\gamma) = A^0 \left[\mathscr{GP}_{\text{mat},1}(\gamma)\, \tilde{\theta}_z^0 - R_1(\gamma)\mathscr{GP}_{\text{mat}}\, \check{\theta}_z^0\right]$$

where $\tilde{\theta}_z^0 = (1, \theta_z^{0\top})^\top$ and $\check{\theta}_z^0 = (0, \theta_z^{0\top})^\top$. Finally, let:

$$\mathscr{E}(\gamma) = C_1^{-1}(\gamma)\mathscr{B}_1(\gamma) - C_2^{-1}(\gamma)\mathscr{B}_2(\gamma)$$

where $\mathscr{B}_2(\gamma) = \mathscr{B} - \mathscr{B}_1(\gamma)$ with $\mathscr{B} = \mathscr{B}_1(\gamma_{\max})$. Let

$$\sigma^2 = \sigma_\epsilon^2 + 2\Sigma_{\epsilon,u}^\top \theta_z^0 + \theta_z^{0\top} \Sigma_u \theta_z^0.$$

With this notation, the null distributions for a LFS are stated below.

Theorem 2.2 (Asymptotic Distributions LFS)**.** *Let* Z *be generated by* (2.1)*,* Y *be generated by* (2.3)*, and* $\hat{Z}$ *be calculated by* (2.4)*. Then under* $\mathbb{H}_0$ *and Assumption 2.1,*
(*i*)

$$\sup_{\gamma \in \Gamma} LR_{T,LFS}^{2SLS}(\gamma) \Rightarrow \sup_{\gamma \in \Gamma} \mathscr{E}^\top(\gamma) \mathscr{Q}^{-1}(\gamma) \mathscr{E}(\gamma),$$

where $Q(\gamma) = \sigma^2 C_1^{-1}(\gamma)\ C\ C_2^{-1}(\gamma)$*;*
(ii)

$$\sup_{\gamma\in\Gamma} W^{2SLS}_{T,LFS}(\gamma) \Rightarrow \sup_{\gamma\in\Gamma} \mathscr{E}^{\top}(\gamma) V^{-1}(\gamma)\mathscr{E}(\gamma),$$

where $V(\gamma)$ *is defined in Definition 2.2 in the Appendix, and, in general,* $V(\gamma) \neq Q(\gamma)$*.*

In both cases, the suprema taken are over $\gamma \in \Gamma$ and this deserves some explanation. For theoretical derivations, it suffices that Γ is a closed interval in the support Γ^0 and that it is bounded away from the end-points of $\Gamma^0 = [\gamma_{min}, \gamma_{max}]$. But in practice, searching over γ includes calculations over the subsamples $\{t : \mathbb{1}_{\{q_t \leq \gamma\}}\}$ and $\{t : \mathbb{1}_{\{q_t > \gamma\}}\}$, which means that the data needs to be sorted into quantiles of q_t. Therefore, in practice, Γ is a set that contains ordered values of q_t encountered in the sample, from a pre-defined lower quantile $\underline{\gamma}$ to predefined upper quantile $\overline{\gamma}$, where $\underline{\gamma} > \gamma_{min}$ and $\overline{\gamma} < \gamma_{max}$. We refer to these upper and lower quantiles as "cut-offs" in the simulation section, and in practice they are chosen so that the subsamples $\{t : \gamma_{min} \leq q_t \leq \underline{\gamma}\}$ and $\{t : \gamma_{max} \geq q_t \geq \overline{\gamma}\}$ are large enough to produce reliable estimates; example cut-offs are the 15% and the 85% quantiles of q_t.

Both asymptotic distributions depend on second moment functionals of the data and the parameters in the FS. But critical values can be calculated by the bootstrap described in Section 2.6.

As shown in Corollary 2.B.1 in the Appendix, the asymptotic distributions remain nonpivotal for both tests even when the errors are conditional homoskedastic. More importantly, because the 2SLS estimators are not conventional, the sup Wald and sup LR tests are in general NOT asymptotically equivalent under conditional homoskedasticity. However, they are equivalent in the just-identified case as shown in Corollary 2.B.1. They are also equivalent in the overidentified case, when x_t and q_t are independent, as stated below and proven in the Appendix.

Corollary 2.1 (to Theorem 2.2)**.** *Let* Z *be generated by* (2.1)*,* Y *be generated by* (2.3)*, and* $\hat{Z}$ *be calculated by* (2.4)*.Then, under* $\mathbb{H}_0$ *and Assumptions 2.1-2.3,*

$$\sup_{\gamma\in\Gamma} LR^{2SLS}_{T,LFS}(\gamma) \Rightarrow \sup_{\lambda\in\Lambda_\epsilon} \frac{\mathscr{BB}_p^{\top}(\lambda)\mathscr{BB}_p(\lambda)}{\lambda(1-\lambda)}, \qquad \sup_{\gamma\in\Gamma} W^{2SLS}_{T,LFS}(\gamma) \Rightarrow \sup_{\lambda\in\Lambda_\epsilon} \frac{\mathscr{BB}_p^{\top}(\lambda)\mathscr{BB}_p(\lambda)}{\lambda(1-\lambda)},$$

where $\mathscr{BB}_p(\lambda) = \mathscr{BM}_p(\lambda) - \lambda\mathscr{BM}_p(1)$*,* $\mathscr{BM}_p(\cdot)$ *is a* $p \times 1$*-vector of independent standard Brownian motions,* $\lambda = \text{Prob}(q_t \leq \gamma)$*,* $\Lambda_\epsilon = [\epsilon_1, 1-\epsilon_2]$*, where* $\epsilon_1 = \text{Prob}(q_t \leq \underline{\gamma})$*,* $\epsilon_2 = \text{Prob}(q_t \leq \overline{\gamma})$*.*

The distribution in Corollary 2.1 is identical that of the sup F and sup Wald break-point tests - see Andrews (1993), Bai and Perron (1998) and Hall et al. (2012) among others. This is due to similarities between threshold and break point models; a break-point model is a special case of a threshold model when $q_t = t/T$.[13] Critical values for these distributions can be found in Andrews

[13]Note, however, that the asymptotics for break-point tests cannot be obtained as a special case of our results here because in general, break-point models are not strictly stationary.

(1993) and Bai and Perron (1998). However, $x_t \perp q_t$ is a case rarely encountered in practice, and we do not consider this case in our simulations.

2.4.4 Asymptotic distributions with a TFS

For this section, we assume that the FS has a threshold ρ^0 (TFS). For stating the asymptotic distributions, similar to A^0 in the previous section, we define

$$A_1^0 = [\Pi_1^0, S^\top]^\top \quad \text{and} \quad A_2^0 = [\Pi_2^0, S^\top]^\top. \tag{2.13}$$

Also, let $a \wedge b = \min(a,b)$ for generic scalars a, b, and define the matrices:

$$C_{A,1}(\gamma) = A_1^0 M_1(\gamma \wedge \rho^0) A_1^{0\top} + A_2^0 \left[M_1(\gamma) - M_1(\gamma \wedge \rho^0)\right] A_2^{0\top}, \tag{2.14}$$

and $C_{A,2} = C_A - C_{A,1}(\gamma)$, where:

$$C_A = C_{A,1}(\gamma_{\max}) = A_1^0 M_1(\rho^0) A_1^{0\top} + A_2^0 M_2(\rho^0) A_2^{0\top},$$

as well as, in line with Section 2.4.3, the "ratios"

$$R_i(\gamma;\rho^0) = M_i(\gamma) M_i^{-1}(\rho^0).$$

The TFS analogs to the LFS processes $B_1(\gamma)$ and $\mathscr{E}(\gamma)$ are defined as:

$$\begin{aligned}\mathscr{B}_{A,1}(\gamma) = {} & A_1^0 \left[\mathscr{G}\mathscr{P}_{\text{mat},1}(\gamma \wedge \rho^0)\tilde{\theta}_z^0 - R_1(\gamma \wedge \rho^0;\rho^0)\mathscr{G}\mathscr{P}_{\text{mat},1}(\rho^0)\check{\theta}_z^0\right] \\ & + A_2^0 \left[\left(\mathscr{G}\mathscr{P}_{\text{mat},1}(\gamma)\tilde{\theta}_z^0 - \mathscr{G}\mathscr{P}_{\text{mat},1}(\gamma \wedge \rho^0)\right)\tilde{\theta}_z^0\right] \\ & - A_2^0 \left[\left(R_2(\gamma \wedge \rho^0;\rho^0) - R_2(\gamma;\rho^0)\right)\mathscr{G}\mathscr{P}_{\text{mat},2}(\rho^0)\check{\theta}_z^0\right].\end{aligned} \tag{2.15}$$

and

$$\mathscr{E}_A(\gamma) = C_{A,1}^{-1}(\gamma)\mathscr{B}_{A,1}(\gamma) - C_{A,2}^{-1}(\gamma)\mathscr{B}_{A,2}(\gamma) \tag{2.16}$$

where

$$\mathscr{B}_{A,2}(\gamma) = \mathscr{B}_A - \mathscr{B}_{A,1}(\gamma)$$

with

$$\mathscr{B}_A = \mathscr{B}_A(\gamma_{\max}) = A_1^0 \mathscr{G}\mathscr{P}_{\text{mat},1}(\rho^0)(\tilde{\theta}_z^0 - \check{\theta}_z^0) + A_2^0 \mathscr{G}\mathscr{P}_{\text{mat},2}(\rho^0)(\tilde{\theta}_z^0 - \check{\theta}_z^0).$$

The more complicated expressions in this case stem from the fact that the relative location of γ and ρ^0 influences the asymptotic distribution of our tests, as Theorem 2.3 shows.

Theorem 2.3 (Asymptotic Distributions TFS)**.** *Let Z be generated by* (2.2)*, Y be generated by* (2.3)*, and $\hat{Z}$ be calculated by* (2.8)*. Under $\mathbb{H}_0$ and Assumptions 2.1 and 2.4,*
(*i*)

$$\sup_{\gamma\in\Gamma} LR_{T,TFS}^{2SLS}(\gamma) \Rightarrow \sup_{\gamma\in\Gamma} \mathscr{E}_A^\top(\gamma) Q_A^{-1}(\gamma) \mathscr{E}_A(\gamma),$$

where $Q_A(\gamma) = \sigma^2 C_{A,1}^{-1}(\gamma)\, C_A\, C_{A,2}^{-1}(\gamma)$*;*
(ii)

$$\sup_{\gamma\in\Gamma} W_{T,TFS}^{2SLS}(\gamma) \Rightarrow \sup_{\gamma\in\Gamma} \mathscr{E}_A^\top(\gamma) V_A^{-1}(\gamma) \mathscr{E}_A(\gamma),$$

where $V_A(\gamma)$ *is defined in Definition 2.3 of the Appendix, and in general,* $V_A(\gamma) \neq Q_A(\gamma)$*.*

Under conditional homoskedasticity, Corollary 2.B.2 in the Appendix shows that, as for a LFS, the sup Wald and sup LR tests are not asymptotically equivalent for a TFS, except for the just identified case $p = q$.

As in Boldea et al. (2017), in this section, the asymptotic distributions are non-pivotal, and don't simplify to the usual break-point distributions expressed in Corollary 2.1. This is not an issue in practice, because critical values can still be obtained by bootstrap, as we discuss in Section 2.6.

2.5 GMM test

In contrast to our paper, Caner and Hansen (2004) propose testing for a threshold using a GMM sup Wald test. To calculate this test, they use the conventional two-step GMM estimators defined in Section 2.3, with estimated variance-covariances:

$$\hat{V}_{i,GMM}(\gamma) = \left(T^{-1} W_i^{\gamma\top} X_i^\gamma \hat{H}_{i,GMM}^{\epsilon^{-1}}(\gamma) X_i^{\gamma\top} W_i^\gamma\right)^{-1}.$$

The Wald test statistic in Caner and Hansen (2004) for $\mathbb{H}_0$ at each γ is:

$$W_T^{\text{GMM}}(\gamma) = T[\hat{\theta}_{1,GMM}^\gamma - \hat{\theta}_{2,GMM}^\gamma]^\top \{\hat{V}_{1,GMM}(\gamma) + \hat{V}_{2,GMM}(\gamma)\}^{-1} [\hat{\theta}_{1,GMM}^\gamma - \hat{\theta}_{2,GMM}^\gamma],$$

and the sup Wald test is $\sup_{\gamma\in\Gamma} W_T^{GMM}(\gamma)$.

For clarity, we reproduce below the asymptotic distribution of this test, which was already derived in Caner and Hansen (2004). Assume that $\mathbb{H}_0$ holds, and let $V_{i,GMM}(\gamma) = \left[N_i(\gamma) H_i^{\epsilon^{-1}}(\gamma) N_i^\top(\gamma)\right]^{-1}$, where $H_i^\epsilon(\gamma)$ is defined in Definition 2.1 of the Appendix. Also, let $N_i(\gamma) = A_i^{0\top} M_i(\gamma)$, and let $\overline{\mathscr{GP}}_1(\gamma)$, be a $q\times 1$ zero mean Gaussian process with covariance kernel equal to $\mathbb{E}[\overline{\mathscr{GP}}_1(\gamma_1)\overline{\mathscr{GP}}_1^\top(\gamma_2)] = H_i^\epsilon(\gamma_1 \wedge \gamma_2)$. Let $\overline{\mathscr{GP}} = \overline{\mathscr{GP}}_1(\gamma_{max})$ and $\overline{\mathscr{GP}}_2(\gamma) = \overline{\mathscr{GP}} - \overline{\mathscr{GP}}_1(\gamma)$.[14] Then Caner and Hansen (2004) show:

Theorem 2.4 (Asymptotic distribution sup Wald GMM)**.** *Let* Z *be generated by* (2.1) *or* (2.2)*, and* Y *be generated by* (2.3)*. Under* $\mathbb{H}_0$ *and Assumptions 2.1 and 2.4,*

$$\begin{aligned}\sup_{\gamma\in\Gamma} W_T^{GMM}(\gamma) \Rightarrow \sup_{\gamma\in\Gamma} &\left[V_{1,GMM}(\gamma) N_1(\gamma) H_1^{\epsilon^{-1}}(\gamma) \overline{\mathscr{GP}}_1(\gamma) - V_{2,GMM}(\gamma) N_2(\gamma) H_2^{\epsilon^{-1}}(\gamma) \overline{\mathscr{GP}}_2(\gamma)\right]^\top \\ &\times \left[V_{1,GMM}(\gamma) + V_{2,GMM}(\gamma)\right]^{-1} \\ &\times \left[V_{1,GMM}(\gamma) N_1(\gamma) H_1^{\epsilon^{-1}}(\gamma) \overline{\mathscr{GP}}_1(\gamma) - V_{2,GMM}(\gamma) N_2(\gamma) H_2^{\epsilon^{-1}}(\gamma) \overline{\mathscr{GP}}_2(\gamma)\right].\end{aligned}$$

[14] In Caner and Hansen (2004), $\overline{\mathscr{GP}} = \lim_{\gamma\to\infty} \overline{\mathscr{GP}}_1(\gamma)$, to account for an unbounded support Γ^0; as discussed before, for all practical purposes, including calculation of critical values, it makes sense to impose $\Gamma^0 = [\gamma_{min}, \gamma_{max}]$, treat $\gamma_{min}, \gamma_{max}$ as fixed values, and therefore define $\overline{\mathscr{GP}} = \overline{\mathscr{GP}}_1(\gamma_{max})$.

The proof is in Caner and Hansen (2004). Theorems 2.2-2.4 show that the 2SLS and GMM tests have different asymptotic distributions in general, but there are two notable exceptions, both for a LFS. First, under conditional homoskedasticity and just identification, a comparison of Corollaries 2.B.1 and 2.B.3 in the Appendix shows that the GMM test distribution looks just like the 2SLS distributions for a LFS, with the difference that the Gaussian processes are generated by ϵ_t rather than $(\epsilon_t + u_t\theta_z^0)$. Second, under Assumptions 2.1-2.3 and a LFS, all the distributions are the same, and identical to the break-point sup F and sup Wald test distributions. This latter result is stated below and proven in the Appendix.

Corollary 2.2 (Corollary to Theorem 2.4)**.** *Let Z be generated by* (2.1) *and Y be generated by* (2.3)*. Then, under $\mathbb{H}_0$, and Assumptions 2.1-2.3,*

$$\sup_{\gamma\in\Gamma} W_T^{GMM}(\gamma) \Rightarrow \sup_{\lambda\in\Lambda_\epsilon} \frac{\mathscr{B}\mathscr{B}_p^\top(\lambda)\mathscr{B}\mathscr{B}_p(\lambda)}{\lambda(1-\lambda)}$$

Note that for a TFS and the same assumptions, the distribution in Corollary 2.2 does not apply.

2.6 Simulations

In this chapter, we investigate the small sample properties of the 2SLS tests and the GMM test. We first introduce the wild fixed-regresssor bootstrap.

2.6.1 Bootstrap and DGP

Bootstrap As shown in Section 2.4, the asymptotic distributions of the proposed test statistics are non-standard and therefore need to be either simulated or bootstrapped.

Simulating the asymptotic distributions involves, for example, simulating the Gaussian processes $\mathscr{E}(\cdot)$ and $\mathscr{E}_A(\cdot)$ in Theorems 2.2-2.4, while keeping x_t, q_t fixed. On the other hand, in simulations, usually $Q(\gamma), V(\gamma), Q_A(\gamma), V_A(\gamma)$ are replaced with consistent estimators based on the initial sample, $\hat{Q}(\gamma), \hat{V}(\gamma), \hat{Q}_A(\gamma), \hat{V}_A(\gamma)$, and are kept fixed across simulations. Using similar arguments to Hansen (1996), Theorem 2, one can show that the critical value simulated in this way converges to the true critical value of the test. However, the randomness of $\hat{Q}(\gamma), \hat{V}(\gamma), \hat{Q}_A(\gamma), \hat{V}_A(\gamma)$ may affect the critical value approximation in finite samples. Therefore, we propose bootstrapping the critical values instead.

Below, we describe the **fixed regressor wild bootstrap** we used for simulating critical values. We first describe it for the 2SLS test and then for the GMM test.

Bootstrap for 2SLS tests:

1. based on the original sample, compute the test statistics in Section 2.4, gathered under the generic name $\hat{G}$:

$$\hat{G} : \sup_{\gamma\in\Gamma} LR^{2SLS}_{T,LFS}(\gamma),\ \sup_{\gamma\in\Gamma} LR^{2SLS}_{T,TFS}(\gamma),\ \sup_{\gamma\in\Gamma} W^{2SLS}_{T,LFS}(\gamma),\ \sup_{\gamma\in\Gamma} LR^{2SLS}_{T,TFS}(\gamma)$$

2. compute the full-sample 2SLS parameter estimates $\hat{\theta} = (\hat{\theta}_z^\top, \hat{\theta}_x^\top)^\top$ for a LFS or for a TFS, using (2.4) or (2.8), and the corresponding residuals for these estimates:

$$\hat{v}_t = (\hat{\epsilon}_t^\top, \hat{u}_t^\top)^\top$$

3. for each bootstrap sample j, draw a random sample $t = 1,\dots,T$ for η_t such that[15]

$$\eta_t = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability } (\sqrt{5}+1)/(2\sqrt{5}) \\ (\sqrt{5}+1)/2 & \text{with probability } (\sqrt{5}-1)/(2\sqrt{5}) \end{cases},$$

   and compute the **wild bootstrap** residuals:

$$\hat{v}_t^{(j)} = \hat{v}_t \eta_t$$

4. keeping x_t, q_t **fixed**, calculate a new bootstrap sample $(y_t^{(j)}, z_t^{(j)})$

$$z_t^{(j)} = \hat{\Pi}^\top x_t + \hat{u}_t^{(j)} \text{ for a LFS or } z_t^{(j)} = \hat{\Pi}_1^\top x_t \mathbb{1}_{\{q_t \le \hat{\rho}\}} + \hat{\Pi}_2^\top x_t \mathbb{1}_{\{q_t > \hat{\rho}\}} + \hat{u}_t^{(j)} \text{ for a TFS}$$
$$y_t^{(j)} = z_t^{(j)\top} \hat{\theta}_z + x_{1t}^\top \hat{\theta}_x + \hat{\epsilon}_t^{(j)}$$

5. using the new sample $(y_t^{(j)}, z_t^{(j)}, x_t, q_t)$ with **fixed regressors** x_t, q_t, recalculate all 2SLS test statistics, gathered under the generic name $\hat{G}^{(j)}$

$$\hat{G}^{(j)} : \sup_{\gamma\in\Gamma} LR^{2SLS,(j)}_{T,LFS}(\gamma),\ \sup_{\gamma\in\Gamma} LR^{2SLS,(j)}_{T,TFS}(\gamma),\ \sup_{\gamma\in\Gamma} W^{2SLS,(j)}_{T,LFS}(\gamma),\ \sup_{\gamma\in\Gamma} W^{2SLS,(j)}_{T,TFS}(\gamma)$$

6. repeat this procedure for $j = 1,\dots,J$ times

7. the 5% bootstrap critical value for each test statistic is equal to the 95% quantile from the empirical distribution $(\hat{G}^{(1)},\dots,\hat{G}^{(J)})$, call it $\hat{G}_{0.95}$

8. if $\hat{G} > \hat{G}_{0.95}$ we reject, else we don't reject.

[15] This distribution for the bootstrap was proposed by Mammen (1993). We also tried the Rademacher-distribution and a standard normal distribution for bootstrapping the residuals. Results do not change by much when using the Rademacher distribution and substantially change for the GMM test when using the standard normal distribution. In particular, this test becomes even more oversized in small samples when using the standard normal distribution. These results are available from the authors upon request.

Bootstrap for the GMM test:

1. based on the original sample, compute the GMM test statistic:

$$\hat{G} = \sup_{\gamma \in \Gamma} W_T^{GMM}(\gamma)$$

2. compute the full-sample two-step GMM parameter estimates $\hat{\theta}_{GMM}$ using the 2SLS estimator $\hat{\theta}$ for a LFS as the first-step GMM estimator; calculate the corresponding residuals:

$$\tilde{\epsilon}_t = y_t - w_t^\top \hat{\theta}_{GMM}$$

3. for each bootstrap sample j, draw a random sample $t = 1, \ldots, T$ for η_t such that[16]

$$\eta_t = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability } (\sqrt{5}+1)/(2\sqrt{5}) \\ (\sqrt{5}+1)/2 & \text{with probability } (\sqrt{5}-1)/(2\sqrt{5}) \end{cases},$$

   and compute the **wild bootstrap** residuals:

$$\tilde{\epsilon}_t^{(j)} = \tilde{\epsilon}_t \eta_t$$

4. keeping z_t, x_t, q_t fixed, calculate a new bootstrap sample $y_t^{(j)}$

$$y_t^{(j)} = w_t^\top \hat{\theta}_{GMM} + \tilde{\epsilon}_t^{(j)}$$

5. using the new sample $(y_t^{(j)}, z_t, x_t, q_t)$ with **fixed regressors** z_t, x_t, q_t, recalculate the GMM test statistic $\hat{G}^{(j)}$

$$\hat{G}^{(j)} = \sup_{\gamma \in \Gamma} W_T^{GMM,(j)}$$

6. the 5% bootstrap critical value for each test statistic is equal to the 95% quantile from the empirical distribution $(\hat{G}^{(1)}, \ldots, \hat{G}^{(J)})$, call it $\hat{G}_{0.95}$

7. if $\hat{G} > \hat{G}_{0.95}$ we reject, else we don't reject.

Our bootstrap is slightly different than the one suggested in Caner and Hansen (2004) for the same test statistic. They suggested setting $y_i^{(j)} = \tilde{\epsilon}_t \eta_t$, therefore computing a "pseudo-sample" that ignores the predictable part of y_t under $\mathbb{H}_0$, which is $(w_t^\top \theta^0)$. Presumably, they do so because the value of θ^0 is irrelevant for the asymptotic distribution of their test statistic. However, θ^0 shows up in the asymptotic distribution of our test statistics, and for the sake of comparison, we

[16] This distribution for the bootstrap was proposed by Mammen (1993). We also tried the Rademacher-distribution and a standard normal distribution for bootstrapping the residuals. Results do not change by much when using the Rademacher distribution and substantially change when using the standard normal distribution. This substantial change is only experienced by the GMM test, which becomes even more oversized in small samples. These results are available from the authors upon request.

compute $y_t^{(j)}$ as suggested in Step 5. Computing $y_t^{(j)}$ as we suggested is a proper wild bootstrap. Compared to Caner and Hansen (2004), it should replicate more closely the sample null behavior of the test.

As we already mentioned in Section 2.3, we investigate two possibilities in case the errors are homoskedastic, namely, a) we know that they are homoskedastic or b) we do not know that they are homoskedastic. In case b) we use the wild bootstrap, as explained above, and the heteroskedasticity robust test statistics, as presented in Sections 2.4 and 2.5. In case a) we make two adjustments to simulate the size and power properties of the tests:

- First, we replace the above wild bootstrap with the fixed regressor $i.i.d.$ bootstrap. That is, we replace step 3 in the wild bootstrap such that $\hat{v}_t^{(j)} \overset{i.i.d.}{\sim} \mathcal{N}(0, \hat{\Sigma}_v)$ with $\hat{\Sigma}_v = T^{-1}\sum_{t=1}^{T} \hat{v}_t \hat{v}_t^\top$ in case of 2SLS. In case of GMM we replace step 3 in the wild bootstrap such that $\tilde{\epsilon}_t^{(j)} \overset{i.i.d.}{\sim} \mathcal{N}(0, \tilde{\sigma}_\epsilon^2)$ with $\tilde{\sigma}_\epsilon^2 = T^{-1}\sum_{t=1}^{T} \tilde{\epsilon}_t^2$.

- Second, we replace second moment functionals which contain v_t by their homoskedasticity analogs. For example, we replace the term $\mathbb{E}[x_t x_t^\top \epsilon_t^2 \mathbb{1}\{q_t \leq \gamma\}]$, which is estimated by $T^{-1}\sum_{t=1}^{T} x_t x_t^\top \hat{\epsilon}_t^2 \mathbb{1}\{q_t \leq \gamma\}$, by its homoskedasticity analog $\sigma_\epsilon^2 \mathbb{E}[x_t x_t^\top \mathbb{1}\{q_t \leq \gamma\}]$, which is estimated by $\hat{\sigma}_\epsilon^2 T^{-1}\sum_{t=1}^{T} x_t x_t^\top \mathbb{1}\{q_t \leq \gamma\}$. We proceed for all other such quantities in the same way. This yields the simplified variance-covariance terms in Corollaries 2.B.1, 2.B.2 and 2.B.3 in Appendix 2.B and consequently simplified sample test statistics to compute.

Empirical Sizes and Size Adjusted Power To calculate the empirical sizes $\hat{\alpha}$ for a nominal significance level α, we repeat the bootstrap procedure MC times, for a certain fixed $\mathbb{H}_0$ DGP but with the original sample redrawn in each simulation draw $s = 1, \ldots, MC$, and set:

$$\hat{\alpha} = \frac{1}{MC}\sum_{s=1}^{MC} \mathbb{1}_{\hat{G}_s > \hat{G}_{0.95,s}}, \tag{2.17}$$

where the subscript s in $\hat{G}_s, \hat{G}_{0.95,s}$ refers to the s^{th} simulated value of $\hat{G}, \hat{G}_{0.95}$. The empirical power is obtained analogously with the DGP under $\mathbb{H}_A$:

$$\hat{\beta} = \frac{1}{MC}\sum_{s=1}^{MC} \mathbb{1}_{\hat{G}_s > \hat{G}_{0.95}}. \tag{2.18}$$

Setting $\hat{G}_{0.95}$ in (2.18) equal to the 95%-quantile of the empirical distribution of a given test statistic yields the size adjusted power.[17]

[17] Note that the size adjusted power is defined/computed such that the considered test has empirical size exactly equal to the required nominal size. This is guaranteed under this setting.

DGP The $\mathbb{H}_0$ DGP used in the simulations for calculating empirical sizes is:

$$y_t = \theta^0_{x_1} + z_t\theta^0_z + \epsilon_t = w_t^\top \theta^0 + \epsilon_t \tag{2.19}$$

$$z_t = (\Pi^0_{1,1} + \Pi^0_{1,2}x_t)\mathbb{1}_{\{q_t \le \rho^0\}} + (\Pi^0_{2,1} + \Pi^0_{2,2}x_t)\mathbb{1}_{\{q_t > \rho^0\}} + u_t \tag{2.20}$$

where $x_t \overset{iid}{\sim} \mathcal{N}(1,1)$, $q_t = x_t + 1$, and x_t, z_t, q_t are scalars. We set:

- $\theta^0_z = \theta^0_{x_1} = 1$.
- $\Pi^0_1 = (\Pi^0_{1,1}, \Pi^0_{1,2})^\top = (1,1)^\top$.
- $\Pi^0_2 = (\Pi^0_{2,1}, \Pi^0_{2,2})^\top = (1,b)^\top$, where we allow $b \in \{0.5, 1, 1.5, 2, 2.5\}$. Note that $b = 1$ corresponds to a LFS, and $b \neq 1$ to a TFS.
- $\rho^0 = 1.75$.

We consider two cases: homoskedasticity and heteroskedasticity. For homoskedasticity, $\epsilon_t = \nu_t$, and for conditional heteroskedasticity, $\epsilon_t = \nu_t \cdot x_t/\sqrt{2}$ with

$$\begin{pmatrix} \nu_t \\ u_t \end{pmatrix} \overset{iid}{\sim} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right). \tag{2.21}$$

We set $J = 500$ and $MC = 1000$.

Note that we chose on purpose a DGP where the parameters in the equation of interest are just-identified rather than over-identified for a LFS. In such a DGP, the GMM estimators are equal to the conventional 2SLS estimators that use the same sub-sample ($\{t : q_t \le \gamma\}$, respectively $\{t : q_t > \gamma\}$) to estimate both the first stage and the second stage (the equation of interest). Therefore, any difference between our LFS tests and the GMM tests should stem from the additional information of a LFS used in the 2SLS tests.

2.6.2 Size

Known Functional Form of the First Stage In this section, for all simulations we know the nature of the FS: LFS or TFS, and we take it as given.

In the case of conditional homoskedastic errors, we present results for the two cases of known and unknown homoskedasticity in Tables 2.1 and 2.2. In the first case of *known* homoskedasticity, the results show that, in small samples, our tests tend to be slightly undersized but stay below the nominal level, while the GMM test is correctly sized or slightly oversized. It seems that the additional FS information does not result in better small sample properties, and the i.i.d. bootstrap correctly replicates all asymptotic distributions. In large samples of about $T = 1000$ all tests are close to their nominal size. However, in the case of *unknown* homoskedasticity, the pattern is entirely different. That is, both of our tests are close or slightly below the

nominal size for all considered sample sizes. This is in strong contrast to the GMM test which is heavily oversized, with empirical sizes of up to 15% for small samples ($T = 100$) and up to 10.3% for large samples ($T = 1000$).

Table 2.1: Empirical sizes for 5% nominal size, a LFS and homoskedastic errors

T	$LR^{2SLS}_{T,LFS}(\gamma)$	$W^{2SLS}_{T,LFS}(\gamma)$	$W^{GMM}_{T}(\gamma)$
Homoskedasticity known			
100	4.7%	3.0%	5.2%
250	4.8%	3.7%	4.0%
500	5.7%	5.3%	4.7%
1000	4.8%	4.8%	4.6%
Homoskedasticity unknown			
100	4.6%	6.5%	15.0%
250	4.6%	5.9%	11.6%
500	5.1%	5.9%	11.2%
1000	5.5%	5.6%	8.5%

Finally, in case of heteroskedastic errors (Tables 2.3 and 2.4), we observe the same pattern as in the case of unknown homoskedasticity. In particular, the GMM Wald-test is severely oversized with empirical sizes of up to 12%. In sharp contrast to this, the 2SLS tests are more adequately sized, with most empirical sizes ranging from about 4.5% to 5.5%, and with the largest empirical size equal to 6.3%.

As we saw in Figure 2.3, for a LFS case, these findings are due to the fact that the wild bootstrap, combined with heteroskedasticity robust test statistics, fails to adequately mimic the empirical distribution of the GMM test for small sample sizes T. Note that there is no systematic difference between the two 2SLS tests, and because they can both be bootstrapped under heteroskedasticity without severe size distortions, we recommend using both.

Unknown Functional Form of the First Stage For a given empirical application, we may not know whether we have a LFS or a TFS. One way to circumvent this issue while avoiding pre-testing or model selection in the FS is to find a misspecification robust functional form for the FS, such as a polynomial approximation. Tables 2.5 and 2.6 presents simulation results for this approach[18]. We find that the empirical sizes of the 2SLS tests are in general too large,

[18]The polynomial approximation was carried out in the following way: First, we simulate the FS as outlined in Section 2.6.1, for both LFS and TFS cases with heteroskedastic errors. Then, we fit a polynomial as an approximation of the FS. We choose the polynomial order by minimizing the associated BIC and keep this order fixed when bootstrapping a given simulated sample. Thus, the polynomial order can vary across simulations. Since a polynomial approximation of the FS is nothing else than having a LFS but with more instruments, we applied the test statistics for a LFS to evaluate these "robustified" tests. We did so for both the case that the true DGP has a LFS, and the case that the true DGP has a TFS. Note that if the true DGP has a LFS, the optimal polynomial order equals 1.

Table 2.2: Empirical sizes for 5% nominal size, a TFS and homoskedastic errors

T	$LR^{2SLS}_{T,TFS}(\gamma)$	$W^{2SLS}_{T,TFS}(\gamma)$	$W^{GMM}_{T}(\gamma)$	$LR^{2SLS}_{T,TFS}(\gamma)$	$W^{2SLS}_{T,TFS}(\gamma)$	$W^{GMM}_{T}(\gamma)$
Homoskedasticty known						
	b=0.5			b=1.5		
100	2.3%	2.8%	4.7%	1.7%	2.2%	5.2%
250	2.9%	3.0%	5.1%	2.4%	2.2%	4.0%
500	2.7%	2.6%	4.6%	3.8%	3.1%	4.5%
1000	4.2%	4.0%	4.2%	3.3%	3.7%	4.3%
	b=2.0			b=2.5		
100	2.6%	2.4%	5.2%	3.2%	1.6%	5.5%
250	3.9%	3.1%	4.9%	4.9%	3.6%	5.1%
500	5.2%	4.1%	4.6%	5.3%	4.0%	4.7%
1000	4.5%	4.9%	4.7%	4.8%	5.0%	4.3%
Homoskedasticty unknown						
	b=0.5			b=1.5		
100	2.4%	6.6%	14.8%	1.6%	4.6%	13.3%
250	1.7%	4.5%	12.4%	2.3%	3.9%	9.4%
500	2.9%	4.8%	13.1%	4.2%	5.1%	9.9%
1000	3.9%	4.6%	10.3%	3.5%	4.8%	7.8%
	b=2.0			b=2.5		
100	2.2%	6.2%	10.9%	2.4%	5.8%	9.7%
250	1.6%	4.1%	8.1%	4.4%	4.9%	7.6%
500	3.1%	4.8%	8.5%	5.4%	6.2%	9.1%
1000	3.9%	4.6%	7.3%	4.7%	5.6%	7.6%

Table 2.3: Empirical sizes for 5% nominal size, a LFS and heteroskedastic errors

T	$LR^{2SLS}_{T,LFS}(\gamma)$	$W^{2SLS}_{T,LFS}(\gamma)$	$W^{GMM}_{T}(\gamma)$
100	5.5%	5.5%	9.8%
250	5.8%	5.0%	9.2%
500	4.8%	5.5%	8.5%
1000	5.2%	5.4%	7.2%

which is not surprising because there is a substantial share of simulations where z_t is not well approximated by a polynomial. The effect of the approximation error reflects more heavily on the 2SLS Wald test, which needs estimates of second moment functionals for the instruments interacted with the threshold variable. When these instruments are, for example, powers of x_t, we need to estimate second moment functionals of powers of x_t. These estimates become increasingly inaccurate as the order of the polynomial increases, leading to higher approximation errors for the 2SLS Wald test than for the 2SLS LR test. Nevertheless, for small samples and a TFS, the 2SLS test is also heavily oversized.

Table 2.4: Empirical sizes for 5% nominal size, a TFS and heteroskedastic errors

T	$LR^{2SLS}_{T,TFS}(\gamma)$	$W^{2SLS}_{T,TFS}(\gamma)$	$W^{GMM}_{T}(\gamma)$	$LR^{2SLS}_{T}(\gamma)$	$W^{2SLS}_{T}(\gamma)$	$W^{GMM}_{T}(\gamma)$
		$b=0.5$			$b=1.5$	
100	5.4%	2.8%	10.5%	4.6%	2.5%	9.1%
250	4.9%	4.4%	10.2%	4.8%	5.1%	8.1%
500	5.2%	3.5%	12.0%	5.5%	4.2%	7.0%
1000	5.6%	4.6%	8.7%	5.2%	4.4%	6.3%
		$b=2.0$			$b=2.5$	
100	4.4%	3.7%	8.7%	5.1%	4.3%	8.5%
250	5.6%	6.0%	7.1%	6.2%	6.3%	6.4%
500	5.9%	4.7%	6.9%	5.7%	4.6%	6.9%
1000	6.0%	5.0%	6.3%	6.0%	5.0%	6.2%

Table 2.5: Empirical Sizes for 2SLS Tests with Polynomial FS Approximation – DGP is LFS

T	$LR^{2SLS}_{T,LFS}$	$W^{2SLS}_{T,LFS}$
100	5.4%	27.3%
250	6.1%	25.8%
500	5.0%	24.8%
1000	5.2%	24.2%

Table 2.6: Empirical Sizes for 2SLS Tests with Polynomial FS Approximation - DGP is TFS

T	$LR^{2SLS}_{T,LFS}(\gamma)$	$W^{2SLS}_{T,LFS}(\gamma)$	$LR^{2SLS}_{T,LFS}(\gamma)$	$W^{2SLS}_{T}(\gamma)$
	$b=0.5$		$b=1.5$	
100	25.4%	4.9%	24.9%	5.9%
250	19.8%	6.2%	19.9%	5.3%
500	16.7%	6.2%	15.4%	6.0%
1000	8.9%	9.2%	9.4%	8.3%
	$b=2.0$		$b=2.5$	
100	16.7%	6.6%	8.3%	8.1%
250	9.2%	8.1%	4.1%	12.5%
500	4.0%	11.8%	2.3%	30.1%
1000	2.1%	28.4%	0.8%	51.8%

Yet another possibility to robustify the FS estimation is to use a TFS and therefore the TFS tests regardless of whether the FS is linear or not. This is sensible since the parameter estimates of a LFS misspecified as a TFS are still consistent, but not efficient. In addition, it is easy to verify that the asymptotic distributions for the TFS case collapse to those of the LFS case. Table 2.7. presents results for this case. Both of our tests are undersized but still relatively close to the true nominal size in this scenario, whereas the GMM test, which does not depend on the FS,

is oversized as discussed in the previous paragraph. Thus, when the researcher does not know whether the FS is linear or a threshold, we recommend using the TFS 2SLS tests. Again, there is no noticeable difference in Table 2.7 among the two tests, so we recommend using both.

Table 2.7: Empirical Sizes for both 2SLS Tests with LFS approximated as a TFS

T	$LR^{2SLS}_{T,TFS}(\gamma)$	$W^{2SLS}_{T,TFS}(\gamma)$
100	3.5%	3.2%
250	2.9%	3.0%
500	3.8%	2.6%
1000	3.7%	3.2%

2.6.3 Power

In this section, we present the size adjusted power of the three tests. We slightly alter the DGP in (2.19) while leaving everything else equal. In particular we set

$$y_t = w_t^\top \theta_1^0 \mathbb{1}_{\{q_t \leq \gamma^0\}} + w_t^\top \theta_2^0 \mathbb{1}_{\{q_t > \gamma^0\}} + \epsilon_t \tag{2.22}$$

with $\theta_1^0 = (1,1)^\top$ as before and $\theta_2^0 = (a,c)^\top$ with $a \in \{1,2\}$, $c \in \{1.25, 1.5, 1.75, 2\}$, and $\delta = c - 1$ the slope threshold size. This allows us to investigate how the power varies with the threshold size, measured by $a-1$ and δ. Finally, we set $\gamma^0 = 2.25$.

We follow Davidson and MacKinnon (1998, Section 6) and plot size-power curves. That is, we plot all possible sizes between 0 and 1 on the x-axis. The sizes used for size-adjusted powers are true empirical sizes in the sense that they are computed based on (simulated) empirical critical values and the empirical distribution function of the test statistics[19]. On the y-axis we plot the size adjusted power which is calculated using the empirical critical values. For reasons of brevity we only plot the cases of a LFS, a TFS with either $b = 0.5$ or $b = 2.5$, sample sizes $T = 250, 1000$ and no change in the EI intercept. Results for the other cases are similar and available upon request.

Figures 2.4 and 2.5 show the result of this exercise when the *true* errors are homoskedastic. In particular, Figure 2.4 plots the size adjusted power when it is known that the errors are homoskedastic and Figure 2.5 when it is unknown. We see that the power of all three tests increases when either the sample size is fixed and the threshold size increases or the threshold size is fixed and the sample size increases. Furthermore, for moderate threshold sizes the tests have very similar power.

If the true errors are indeed heteroskedastic, as is the case in Figure 2.6, then we observe the same pattern as in the homoskedastic case.

[19] The empirical critical values are computed under the DGP of Section 2.6.2. Of course, other $\mathbb{H}_0$-DGPs are possible (e.g. averaging over θ_1^0 and θ_2^0) but it seems natural to take that of Section 2.6.2 for easy comparison.

Figure 2.4: Size-adjusted power curves - known homoskedasticity

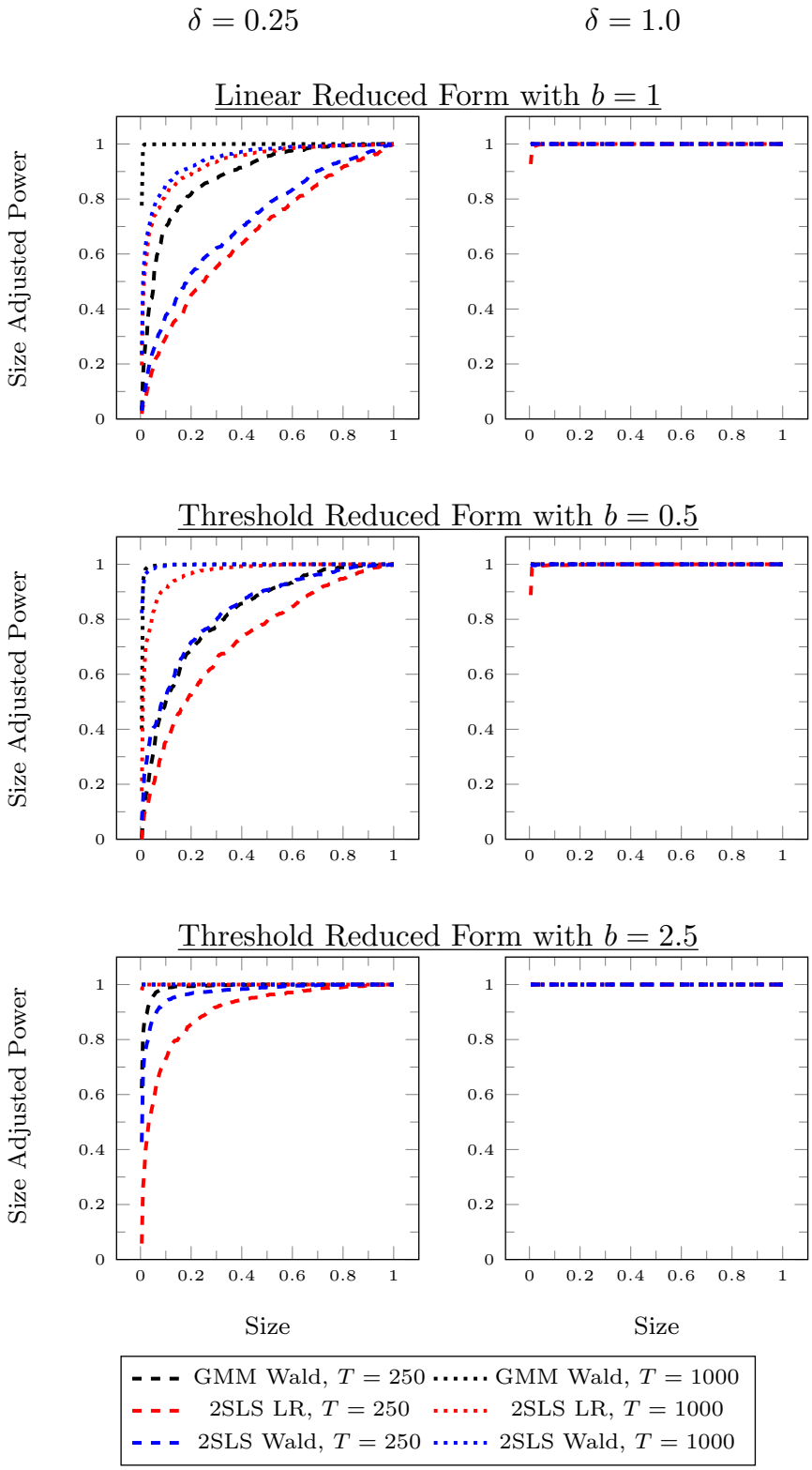


Figure 2.5: Size-adjusted power curves - unknown homoskedasticity

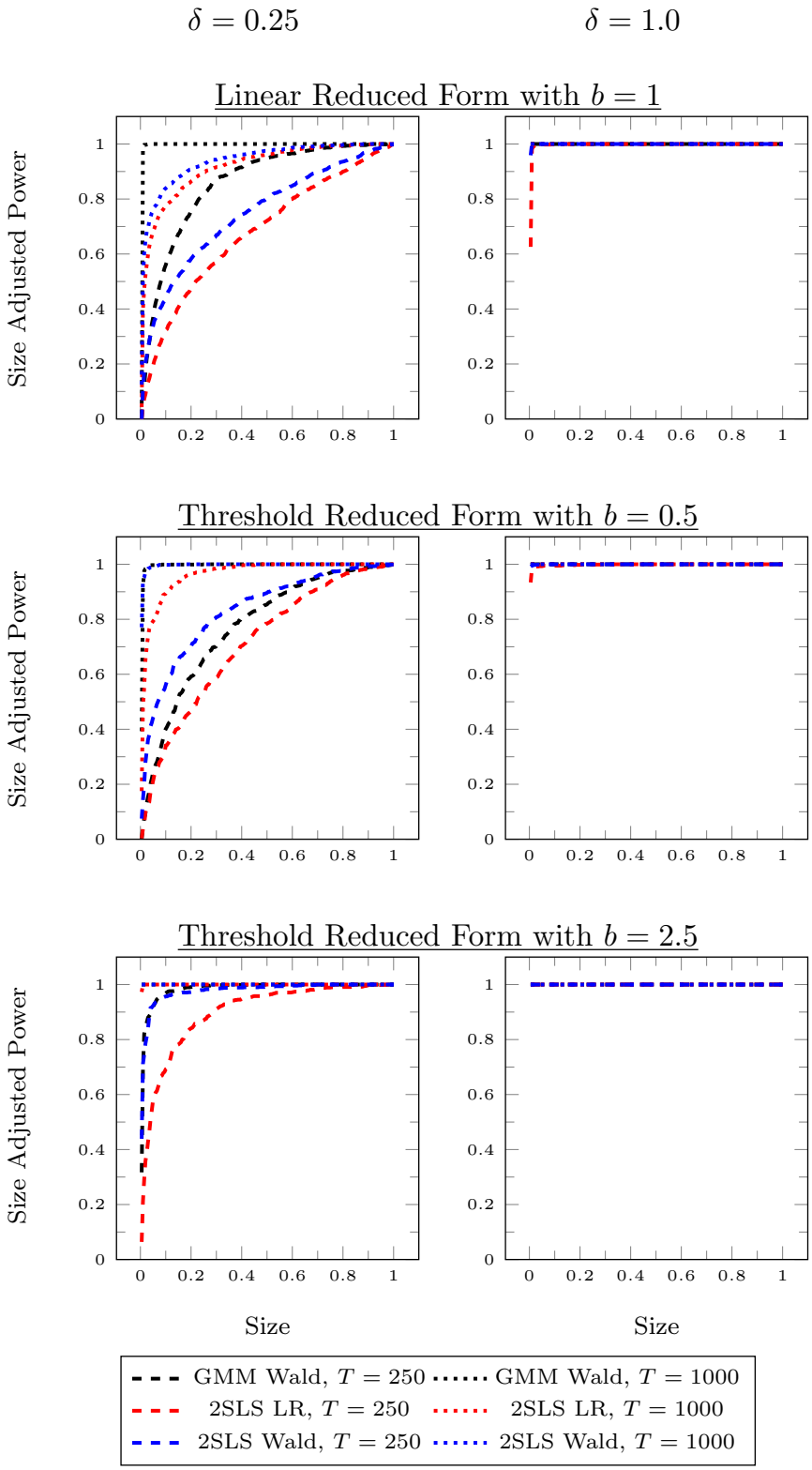
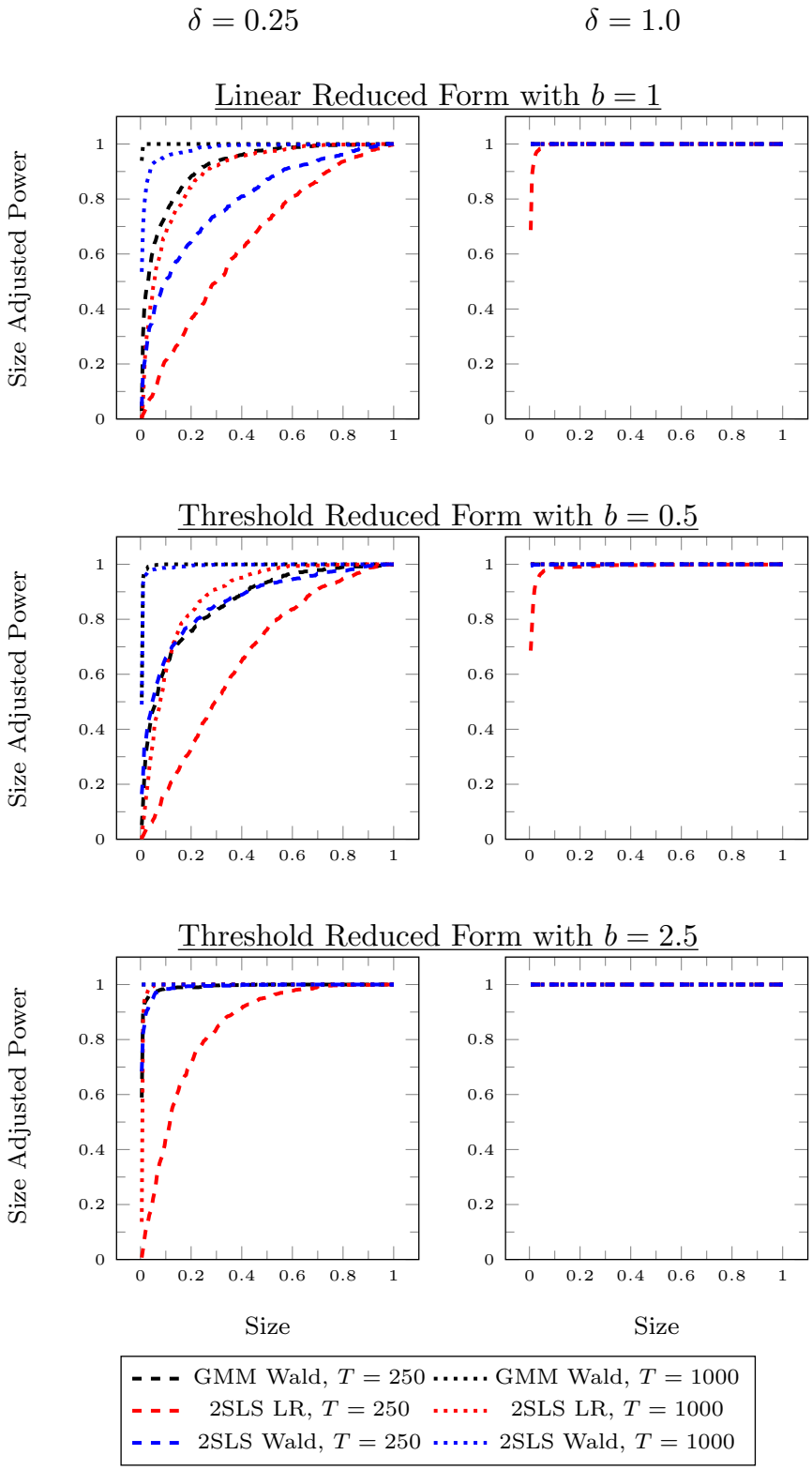


Figure 2.6: Size-adjusted power curves - heteroskedasticity



However, there are is also an interesting difference for small threshold size and/or small sample size: across most cases, the sup Wald tests outperform the sup LR test. Of course, this difference vanishes as the sample size and/or the threshold size increases. Moreover, it seems that the GMM sup Wald test has better power than the 2SLS sup Wald but only in the case of small samples and small threshold values.

Even though our simulations indicate that the sup Wald tests are better than the 2SLS sup LR test in terms of power, we know from Caner and Hansen (2004) that under the alternative, the γ at which the supremum is obtained for the sup LR test is a consistent threshold estimator whether we have an LFS or a TFS, so it is useful to compute the 2SLS sup LR test as well.

2.7 Conclusion

In this paper, we propose two novel threshold tests for linear models with endogenous regressors, a sup LR and a sup Wald test. These tests are based on 2SLS estimation and explicitly account for a possible threshold effect in the FS. We derive the asymptotic distributions of our tests, which are non-pivotal but whose critical values or p-values can easily be bootstrapped. Our simulation study shows that both tests behaves well in small samples, and their size and power compare favorably to an existing GMM based sup Wald test. We therefore recommend using both when testing for a threshold.

Appendix 2.A Definitions

Definition 2.1 (H **and** $\hat{H}$ **matrices).**

$$H_1^u(\gamma) = \mathbb{E}[x_t x_t^\top (u_t^\top \theta_z^0)^2 \mathbb{1}_{\{q_t \leq \gamma\}}] \qquad H_2^u(\gamma) = \mathbb{E}[x_t x_t^\top (u_t^\top \theta_z^0)^2 \mathbb{1}_{\{q_t > \gamma\}}]$$
$$H_1^\epsilon(\gamma) = \mathbb{E}[x_t x_t^\top \epsilon_t^2 \mathbb{1}_{\{q_t \leq \gamma\}}] \qquad H_2^\epsilon(\gamma) = \mathbb{E}[x_t x_t^\top \epsilon_t^2 \mathbb{1}_{\{q_t > \gamma\}}]$$
$$H_1^{\epsilon,u}(\gamma) = \mathbb{E}[x_t x_t^\top \epsilon_t u_t^\top \theta_z^0 \mathbb{1}_{\{q_t \leq \gamma\}}] \qquad H_2^{\epsilon,u}(\gamma) = \mathbb{E}[x_t x_t^\top \epsilon_t u_t^\top \theta_z^0 \mathbb{1}_{\{q_t > \gamma\}}]$$
$$H_1(\gamma) = H_1^u(\gamma) + 2H_1^{\epsilon,u}(\gamma) + H_1^u(\gamma) \qquad H_2(\gamma) = H_2^u(\gamma) + 2H_2^{\epsilon,u}(\gamma) + H_2^u(\gamma).$$

Also, let $H = H_1(\gamma_{\max}) = \mathbb{E}[x_t x_t^\top (\epsilon_t + u_t^\top \theta_z^0)^2]$ *and* $H^u = H_1^u(\gamma_{\max}) = \mathbb{E}[x_t x_t^\top (u_t^\top \theta_z^0)^2]$.

Their estimators are constructed under $\mathbb{H}_0$. *Let* $\hat{z}_t$ *and therefore* $\hat{w}_t = (\hat{z}_t^\top, x_{1t}^\top)^\top$ *be calculated by* (2.4) *for a LFS and by* (2.8) *for a TFS. Let* $\hat{u}_t = z_t - \hat{z}_t$ *and* $\hat{\epsilon}_t = y_t - w_t'\hat{\theta}$, *where* $\hat{\theta} = (\hat{W}^\top \hat{W})^{-1} \hat{W}^\top Y$, *the full sample 2SLS estimator, partitioned as* $\hat{\theta} = (\hat{\theta}_z^\top, \hat{\theta}_x^\top)^\top$. *The sample analogs of all H matrices above are denoted with a hat accent* $\hat{H}$, *and replace* $\mathbb{E}$ *with* $T^{-1}\sum_{t=1}^T$, *and* $\epsilon_t, u_t, \theta_z^0$ *with* $\hat{\epsilon}_t, \hat{u}_t, \hat{\theta}_z$; *for example,* $\hat{H}_1^\epsilon(\gamma) = T^{-1}\sum_{t=1}^T x_t x_t^\top \hat{\epsilon}_t^2 \mathbb{1}_{\{q_t \leq \gamma\}}$.

Definition 2.2 ($V(\gamma)$ **and** $\hat{V}(\gamma)$)**.** *We have a LFS as in* (2.1)*. Then:*

$$
\begin{aligned}
V(\gamma) &= V_1(\gamma) + V_2(\gamma) - V_{12}(\gamma) - V_{12}^{\top}(\gamma) \\
V_i(\gamma) &= C_i^{-1}(\gamma) A^0 \Big[H_i(\gamma) + R_i(\gamma) H^u R_i^{\top}(\gamma) - [H_i^{\epsilon,u}(\gamma) + H_i^u(\gamma)] R_i^{\top}(\gamma) \\
&\qquad - R_i(\gamma) [H_i^{\epsilon,u}(\gamma) + H_i^u(\gamma)] \Big] A^{0\top} C_i^{-1}(\gamma), \quad i = 1,2 \\
V_{12}(\gamma) &= -C_1^{-1}(\gamma) A^0 \Big[[H_1^{\epsilon,u}(\gamma) + H_1^u(\gamma)] R_2^{\top}(\gamma) + R_1(\gamma) [H_2^{\epsilon,u}(\gamma) + H_2^u(\gamma)] \\
&\qquad - R_1(\gamma) H^u R_2^{\top}(\gamma) \Big] A^{0\top} C_2^{-1}(\gamma).
\end{aligned}
$$

$V_A(\gamma)$ is constructed by replacing all quantities in the definition of $V_A(\gamma)$ by their sample analogs, denoted with a hat accent. For example, $\hat{V}_i(\gamma) = \hat{C}_i^{-1}(\gamma)\hat{A}\Big[\hat{H}_i(\gamma) + \hat{R}_i(\gamma)\hat{H}^u\hat{R}_i^{\top}(\gamma) - [\hat{H}_i^{\epsilon,u}(\gamma) + \hat{H}_i^u(\gamma)]\hat{R}_i^{\top}(\gamma) - \hat{R}_i(\gamma)[\hat{H}_i^{\epsilon,u}(\gamma) + \hat{H}_i^u(\gamma)]\Big]\hat{A}^{\top}\hat{C}_i^{-1}(\gamma)$, with $\hat{A} = [\hat{\Pi}, S^{\top}]^{\top}$, $\hat{C}_i(\gamma) = \hat{A}\hat{M}_i(\gamma)\hat{A}^{\top}$, $\hat{M}_1(\gamma) = T^{-1}\sum_{t=1}^{T} x_t x_t^{\top} \mathbb{1}_{\{q_t \le \gamma\}}$,
$\hat{M}_2(\gamma) = T^{-1}\sum_{t=1}^{T} x_t x_t^{\top} \mathbb{1}_{\{q_t > \gamma\}}$, $\hat{M} = \hat{M}_1(\gamma_{max})$, $\hat{R}_i(\gamma) = \hat{M}_i(\gamma)\hat{M}^{-1}$.

Definition 2.3 ($V_A(\gamma)$ **and** $\hat{V}_A(\gamma)$)**.** *We have a TFS as in* (2.2)*. Then:*

$$
\begin{aligned}
V_A(\gamma) &= V_{A,1}(\gamma) + V_{A,2}(\gamma) - V_{A,12}(\gamma) - V_{A,12}^{\top}(\gamma) \\
V_{A,1}(\gamma) &= C_{A,1}^{-1}(\gamma) A_1^0 \Big[H_1(\gamma) + R_1(\gamma;\rho^0) H_1^u(\rho^0) R_1^{\top}(\gamma;\rho^0) - [H_1^{\epsilon,u}(\gamma) + H_1^u(\gamma)] R_1^{\top}(\gamma;\rho^0) \\
&\qquad - R_1(\gamma;\rho^0) [H^{\epsilon,u}(\gamma) + H_1^u(\gamma)] \Big] A_1^{0\top} C_{A,1}^{-1}(\gamma) \\
V_{A,2}(\gamma) &= C_{A,2}^{-1}(\gamma) \Big[A_2^0 \; H_2^{\epsilon}(\rho^0) \; A_2^0 + A_1^0 \; [H_1^{\epsilon}(\rho^0) - H_1^{\epsilon}(\gamma) + H_1^u(\gamma) + R_1(\gamma;\rho^0) H_1^u(\rho^0) R_1^{\top}(\gamma;\rho^0) \\
&\quad + R_1(\gamma;\rho^0) [H_1^{\epsilon,u}(\rho^0) - H_1^{\epsilon,u}(\gamma) - H_1^u(\gamma)] \\
&\quad + [H_1^{\epsilon,u}(\rho^0) - H_1^{\epsilon,u}(\gamma) - H_1^u(\gamma)] R_1^{\top}(\gamma;\rho^0)] \; A_1^{0\top} \Big] C_{A,2}^{-1}(\gamma) \\
V_{A,12}(\gamma) &= -C_{A,1}^{-1}(\gamma) A_1^0 \; [H_1^u(\gamma) + H_1^{\epsilon,u}(\gamma) + R_1(\gamma;\rho^0)(H_1^{\epsilon,u}(\rho^0) - H_1^{\epsilon,u}(\gamma) - H_1^u(\gamma)) \\
&\quad - (H_1^{\epsilon,u}(\gamma) + H_1^u(\gamma)) R_1^{\top}(\gamma;\rho^0) \\
&\quad + R_1(\gamma;\rho^0) H_1^u(\rho^0) R_1^{\top}(\gamma;\rho^0)] \; A_1^{0\top} C_{A,2}^{-1}(\gamma)
\end{aligned}
$$

whenever $\gamma \le \rho^0$. When $\gamma > \rho^0$, then

$$
\begin{aligned}
V_{A,1}(\gamma) = C_{A,1}^{-1}(\gamma) \Big[& A_1^0 H_1^{\epsilon}(\rho^0) A_1^{0\top} + A_2^0 H_2^{\epsilon}(\rho^0) A_2^{0\top} + A_2^0 [H_2^u(\gamma) - H_2^{\epsilon}(\gamma)] A_2^{0\top} \\
& + A_2^0 R_2(\gamma;\rho^0) H_2^u(\rho^0) R_2^{\top}(\gamma;\rho^0) A_2^{0\top} \\
& + A_2^0 H_2^{\epsilon,u}(\rho^0) R_2^{\top}(\gamma;\rho^0) A_2^{0\top} \\
& + A_2^0 R_2(\gamma;\rho^0) H_2^{\epsilon,u}(\rho^0) A_2^{0\top} \\
& - A_2^0 [H_2^{\epsilon,u}(\gamma) + H_2^u(\gamma)] R_2^{\top}(\gamma;\rho^0) A_2^{0\top} \\
& - A_2^0 R_2(\gamma;\rho^0) [H_2^{\epsilon,u}(\gamma) + H_2^u(\gamma)] A_2^{0\top} \Big] C_{A,1}^{-1}(\gamma)
\end{aligned}
$$

$$
\begin{aligned}
V_{A,2}(\gamma) = C_{A,2}^{-1}(\gamma) A_2^0 \Big[& H_2(\gamma) + R_2(\gamma;\rho^0) H_2^u(\rho^0) R_2^\top(\gamma;\rho^0) \\
& - [H_2^{\epsilon,u}(\gamma) + H_2^u(\gamma)] R_2^\top(\gamma;\rho^0) \\
& - R_2(\gamma;\rho^0)[H_2^{\epsilon,u}(\gamma) + H_2^u(\gamma)] \Big] A_2^{0\top} C_{A,2}^{-1}(\gamma) \\
V_{A,12}(\gamma) = -C_{A,1}^{-1}(\gamma) A_2^0 \Big[& [H_2^{\epsilon,u}(\gamma) + H_2^u(\gamma)] + H_2^{\epsilon,u}(\rho^0) R_2^\top(\gamma;\rho^0) \\
& + R_2(\gamma;\rho^0) H_2^u(\rho^0) R_2^\top(\gamma;\rho^0) \\
& - [H_2^{\epsilon,u}(\gamma) + H_2^u(\gamma)] R_2^\top(\gamma;\rho^0) \\
& - R_2(\gamma;\rho^0)[H_2^{\epsilon,u}(\gamma) + H_2^u(\gamma)] \Big] A_2^{0\top} C_{A,2}^{-1}(\gamma).
\end{aligned}
$$

$\hat{V}_A(\gamma)$ is constructed by replacing all quantities in the definition of $V_A(\gamma)$ by their sample analogs, denoted with a hat accent. For example, $\hat{C}_{A,1} = \hat{A}_1 \hat{M}_1(\gamma \wedge \rho) \hat{A}_1^\top + \hat{A}_2[\hat{M}_1(\gamma) - \hat{M}_1(\gamma \wedge \rho)]\hat{A}_2^\top$, $\hat{A}_i = [\hat{\Pi}_i, S^\top]^\top$ and $\hat{R}_i(\gamma;\hat{\rho}) = \hat{M}_i(\gamma)\hat{M}_i^{-1}(\hat{\rho})$.

Appendix 2.B Proofs

In what follows, we use the symbol K to denote a strictly positive constant. Whenever needed, we use a subscript to distinguish among different constants.

For any $m \times 1$-vector x we denote by $\|x\|_2 = \sqrt{\sum_{i=1}^m x_i^2}$ the Euclidean norm. Moreover, for any real $m \times n$-matrix X we denote by $\|X\|_F = \sqrt{\text{tr}(X^\top X)} = \sqrt{\text{tr}(XX^\top)}$ the Frobenius matrix-norm which is submultiplicative, i.e. for two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times l}$ it holds that $\|AB\|_F \leq \|A\|_F \|B\|_F$, and is compatible with the Euclidean norm, i.e. for a matrix $A \in \mathbb{R}^{m\times n}$ and a vector $x \in \mathbb{R}^{n\times 1}$ it holds that $\|Ax\|_2 \leq \|A\|_F \|x\|_2$. Also note that, for two vectors $u, v \in \mathbb{R}^{n\times 1}$ it holds that $\|uv^\top\|_F = \sqrt{\sum_i \sum_j |u_i v_j|^2} = \sqrt{\sum_i |u_i|^2 \sum_j |v_j|^2} = \sqrt{\sum_i |u_i|^2}\sqrt{\sum_j |v_j|^2} = \|u\|_2 \cdot \|v\|_2$. Furthermore, we denote by I_m the $m \times m$-identity matrix and by $\mathbf{0}_{m\times n}$ an $m \times n$-matrix of zeros.

To simplify notation, we define the following sets $\mathcal{T}_1(\gamma) = \{t : \mathbb{1}_{\{q_t \leq \gamma\}}\}$ and $\mathcal{T}_2(\gamma) = \{: \mathbb{1}_{\{q_t > \gamma\}}\}$. These sets partition the data according to the decision rules $\mathbb{1}_{\{q_t \leq \gamma\}}$ and $\mathbb{1}_{\{q_t > \gamma\}}$, respectively, and will be convenient to display sums.
Moreover, we define $\tilde{\epsilon} = \epsilon + (Z - \hat{Z})\theta_z^0$ and $s = \epsilon + u\theta_z^0$. Also, let $\bar{u}_t = \text{vec}(u_t^\top, \mathbf{0}_{1\times p_2})^\top$ denote the augmented FS error. This way, we can write $w_t = A^0 x_t + \bar{u}_t$ for a LFS. Note that $\tilde{\epsilon}$ can also be partitioned into regimes, with $\tilde{\epsilon}_1^\gamma = \epsilon_1^\gamma + (Z - \hat{Z})_1^\gamma \theta_z^0$ and $\tilde{\epsilon}_2^\gamma = \epsilon_2^\gamma + (Z - \hat{Z})_2^\gamma \theta_z^0$.

All convergence results, if not otherwise stated, are uniformly in γ. Moreover, $\xrightarrow{p}$ denotes convergence in probability and $\Rightarrow$ denotes weak convergence in the Skorokhod-metric.

Proofs for Section 2.4.3: 2SLS tests and a LFS

To prove Theorem 2.2, we first provide four Lemmata and their proofs.

Lemma 2.B.1. *Suppose Assumption 2.1 holds. Then*

$$T^{-1/2}\operatorname{vec}(X_1^{\gamma\top}v) \Rightarrow \mathscr{GP}_1(\gamma)$$

where $\mathscr{GP}_1(\gamma)$ is a zero-mean Gaussian Process with covariance function

$$\mathscr{C}_{\mathscr{GP}}(\gamma_1,\gamma_2) = \mathbb{E}[\mathscr{GP}_1(\gamma_1)\mathscr{GP}_1^\top(\gamma_2)] = \mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t\le(\gamma_1\wedge\gamma_2)\}}]$$

Proof of Lemma 2.B.1. Recall that X is a $T\times q$-matrix and v is a $T\times(1+p_1)$-matrix, both satisfying Assumption 1. Further, let $v_{:,i}$ denote the i-th column of the matrix v. Then, by Hansen (1996, Theorem 1)

$$T^{-1/2}X_1^{\gamma\top}v_{:,i} \Rightarrow \mathscr{GP}_1^i(\gamma)$$

and therefore

$$T^{-1/2}\operatorname{vec}(X_1^{\gamma\top}v) \Rightarrow \begin{pmatrix} \mathscr{GP}_1^1(\gamma) \\ \vdots \\ \mathscr{GP}_1^{1+p_1}(\gamma) \end{pmatrix}. \tag{2.23}$$

By Hansen (1996, Theorem 1), $\mathscr{GP}_1^i(\gamma)$ is a zero-mean Gaussian Process with covariance function

$$\mathscr{C}^i_{\mathscr{GP}}(\gamma_1,\gamma_2) = \mathbb{E}[x_t x_t^\top v_{i,t}^2\mathbb{1}_{\{q_t\le(\gamma_1\wedge\gamma_2)\}}]. \tag{2.24}$$

Similarly, it holds that

$$\mathscr{C}^{i,j}_{\mathscr{GP}}(\gamma_1,\gamma_2) = \mathbb{E}[\mathscr{GP}_1^i(\gamma_1)\mathscr{GP}_1^{j\top}(\gamma_2)] = \mathbb{E}[x_t x_t^\top v_{i,t}v_{j,t}\mathbb{1}_{\{q_t\le(\gamma_1\wedge\gamma_2)\}}]. \tag{2.25}$$

Combining (2.24) and (2.25),

$$\mathscr{C}_{\mathscr{GP}}(\gamma_1,\gamma_2) = \mathbb{E}[\mathscr{GP}_1(\gamma_1)\mathscr{GP}_1^\top(\gamma_2)] = \mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t\le(\gamma_1\wedge\gamma_2)\}}]. \tag{2.26}$$

Results (2.23) and (2.26) complete the proof. □

Lemma 2.B.2. *Suppose Assumption 2.1 holds. Then*

(*i*) $T^{-1}\hat{W}_1^{\gamma\top}\hat{W}_1^\gamma \xrightarrow{p} A^0 M_1(\gamma)A^{0\top} \equiv C_1(\gamma)$

(*ii*) $T^{-1/2}\hat{W}_1^{\gamma\top}\tilde{\epsilon}_1^\gamma \Rightarrow A^0\left(\mathscr{GP}_{\text{mat},1}(\gamma)\tilde{\theta}_z^0 - M_1(\gamma)M^{-1}\mathscr{GP}_{\text{mat},1}\breve{\theta}_z^0\right)$.

Proof of Lemma 2.B.2. First, we prove claim (*i*) and then claim (*ii*).
Claim (*i*)**:** The FS predicted values are

$$\hat{Z} = X\hat{\Pi} \tag{2.27}$$

and

$$T^{1/2}(\hat{\Pi}-\Pi^0) = \left(T^{-1}X^\top X\right)^{-1}\left(T^{-1/2}X^\top u\right). \tag{2.28}$$

By Hansen (1996, Theorem 1), it holds uniformly in γ that

$$T^{-1}X_1^{\gamma\top}X_1^{\gamma} \xrightarrow{a.s.} M_1(\gamma), \text{ and } T^{-1}X^{\top}X \xrightarrow{a.s.} M. \tag{2.29}$$

This implies that $T^{-1}X^{\top}X = \mathcal{O}_p(1)$. By Lemma 2.B.1, $T^{-1/2}X^{\top}u = \mathcal{O}_p(1)$. Therefore, $T^{1/2}(\hat{\Pi} - \Pi^0) = \mathcal{O}_p(1)$ and so $\hat{\Pi} - \Pi^0 = o_p(1)$. Therefore, uniformly in γ,

$$T^{-1}\hat{Z}_1^{\gamma\top}\hat{Z}_1^{\gamma} = \hat{\Pi}^{\top}\left(T^{-1}X_1^{\gamma\top}X_1^{\gamma}\right)\hat{\Pi} \xrightarrow{p} \Pi^{0\top}M_1(\gamma)\Pi^0. \tag{2.30}$$

Last, with S the selection matrix such that $x_{1t} = x_t S$, it holds that

$$\hat{W}_1^{\gamma} = \begin{bmatrix}\hat{Z}_1^{\gamma} & X_{1,1}^{\gamma}\end{bmatrix} = \begin{bmatrix}X_1^{\gamma}\hat{\Pi} & X_{1,1}^{\gamma}\end{bmatrix} = X_1^{\gamma}\begin{bmatrix}\hat{\Pi} & S\end{bmatrix} = X_1^{\gamma}\hat{A}^{\top}. \tag{2.31}$$

Therefore, by (2.30) and (2.31) and uniformly in γ,

$$T^{-1}\hat{W}_1^{\gamma\top}\hat{W}_1^{\gamma} = \hat{A}\left(T^{-1}X_1^{\gamma\top}X_1^{\gamma}\right)\hat{A}^{\top} \xrightarrow{p} A^0 M_1(\gamma)A^{0\top} \equiv C_1(\gamma).$$

Claim (ii)**:** By (2.27) it follows that

$$T^{-1/2}\hat{Z}_1^{\gamma\top}\tilde{\epsilon}_1^{\gamma} = \hat{\Pi}^{\top}\big(\underbrace{T^{-1/2}X_1^{\gamma\top}(\epsilon_1^{\gamma} + u_1^{\gamma}\theta_z^0)}_{=\text{(I)}} - \underbrace{T^{-1/2}X_1^{\gamma\top}X_1^{\gamma}(\hat{\Pi} - \Pi^0)\theta_z^0}_{=\text{(II)}}\big). \tag{2.32}$$

Next, we analyze the limiting behavior of (I) and (II). Recalling that $\tilde{\theta}_z^0 = (1, \theta_z^{0\top})^{\top}$,

$$I = T^{-1/2}X_1^{\gamma\top}(\epsilon_1^{\gamma} + u_1^{\gamma}\theta_z^0) = T^{-1/2}[X_1^{\gamma\top}\epsilon_1^{\gamma}, X_1^{\gamma\top}u_1^{\gamma}]\tilde{\theta}_z^0$$

and thus, by Lemma 2.B.1, uniformly in γ:

$$T^{-1/2}[X_1^{\gamma\top}\epsilon_1^{\gamma}, X_1^{\gamma\top}u_1^{\gamma}]\tilde{\theta}_z^0 \Rightarrow \mathcal{G}\mathcal{P}_{\text{mat},1}(\gamma)\tilde{\theta}_z^0. \tag{2.33}$$

By (2.28), term (II) in (2.32) satisfies

$$II = T^{-1/2}X_1^{\gamma\top}X_1^{\gamma}(\hat{\Pi} - \Pi^0)\theta_z^0 = \left(T^{-1}X_1^{\gamma\top}X_1^{\gamma}\right)\left(T^{-1}X^{\top}X\right)^{-1}\left(T^{-1/2}X^{\top}u\theta_z^0\right). \tag{2.34}$$

Recalling that $\check{\theta}_z^0 = (0, \theta_z^{0\top})^{\top}$,

$$T^{-1/2}X^{\top}u\theta_z^0 = T^{-1/2}X^{\top}\epsilon \cdot 0 + T^{-1/2}X^{\top}u\theta_z^0 = T^{-1/2}[X^{\top}\epsilon, X^{\top}u]\check{\theta}_z^0 \tag{2.35}$$

So, by (2.29), (2.34)–(2.35) and Lemma 2.B.1, uniformly in γ,

$$T^{-1/2}X_1^{\gamma\top}X_1^{\gamma}(\hat{\Pi} - \Pi^0)\theta_z^0 \Rightarrow M_1(\gamma)M^{-1}\mathcal{G}\mathcal{P}_{\text{mat},1}\check{\theta}_z^0. \tag{2.36}$$

Next, because for any $a, b = \mathcal{O}p(1)$, $\hat{\Pi}^{\top}(a - b) = \Pi^{0\top}(a - b) + o_p(1)$, (2.33) and (2.36) together with (2.32) yield, uniformly in γ,

$$T^{-1/2}\hat{Z}_1^{\gamma\top}\tilde{\epsilon}_1^{\gamma} \Rightarrow \Pi^{0\top}\left(\mathcal{G}\mathcal{P}_{\text{mat},1}(\gamma)\tilde{\theta}_z^0 - M_1(\gamma)M^{-1}\mathcal{G}\mathcal{P}_{\text{mat},1}\check{\theta}_z^0\right). \tag{2.37}$$

Last, because $\hat{W}_1^{\gamma\top} = \begin{bmatrix}\hat{Z}_1^{\gamma} & X_{1,1}^{\gamma}\end{bmatrix} = X_1^{\gamma}\hat{A}^{\top}$ (see (2.31)) it immediately follows with (2.37) that, uniformly in γ,

$$T^{-1/2}\hat{W}_1^{\gamma\top}\tilde{\epsilon}_1^{\gamma} \Rightarrow \mathcal{B}_1(\gamma), \tag{2.38}$$

proving claim (ii). □

Lemma 2.B.3. *Suppose Assumption 2.1 holds and define* $\hat{\theta}^\gamma = \text{vec}(\hat{\theta}_1^\gamma, \hat{\theta}_2^\gamma)$*, and* $\bar{\theta}^0 = \text{vec}(\theta^0, \theta^0)$*. Then, under* $\mathbb{H}_0$ *and for a fixed* γ*:*

$$T^{1/2}(\hat{\theta}^\gamma - \bar{\theta}^0) \Rightarrow \mathcal{N}(0, \Sigma^\gamma)$$

with

$$\Sigma^\gamma = \begin{bmatrix} V_1(\gamma) & V_{12}(\gamma) \\ V_{12}^\top(\gamma) & V_2(\gamma), \end{bmatrix}$$

where $V_1(\gamma), V_2(\gamma)$ *and* $V_{12}(\gamma)$ *are defined in Definition 2.2.*

Proof of Lemma 2.B.3. First, we define the following quantities

$$\bar{W} = \begin{bmatrix} \hat{W}_1^\gamma & \mathbf{0} \\ \mathbf{0} & \hat{W}_2^\gamma \end{bmatrix}, \bar{Y} = \begin{bmatrix} Y_1^\gamma \\ Y_2^\gamma \end{bmatrix}, \hat{\theta}^\gamma = \begin{bmatrix} \hat{\theta}_1^\gamma \\ \hat{\theta}_2^\gamma \end{bmatrix}.$$

Thus, the 2SLS estimator is given by

$$\hat{\theta}^\gamma = (\bar{W}^\top \bar{W})^{-1} \bar{W}^\top \bar{Y} = \bar{\theta}^0 + (\bar{W}^\top \bar{W})^{-1} \bar{W}^\top \bar{\tilde{\epsilon}}.$$

where

$$\bar{\tilde{\epsilon}} = \begin{bmatrix} \tilde{\epsilon}_1^\gamma \\ \tilde{\epsilon}_2^\gamma \end{bmatrix} = \begin{bmatrix} \epsilon_1^\gamma + (Z - \hat{Z})_1^\gamma \theta_z^0 \\ \epsilon_2^\gamma + (Z - \hat{Z})_2^\gamma \theta_z^0 \end{bmatrix}.$$

By Lemma 2.B.2,

$$T^{1/2}(\hat{\theta}^\gamma - \bar{\theta}^0) \Rightarrow \begin{bmatrix} C_1^{-1}(\gamma)\mathscr{B}_1(\gamma) \\ C_2^{-1}(\gamma)\mathscr{B}_2(\gamma) \end{bmatrix}.$$

Thus, we are left to derive

$$\Sigma^\gamma = \begin{bmatrix} \text{Var}[C_1^{-1}(\gamma)\mathscr{B}_1(\gamma)] & \text{Cov}[C_1^{-1}(\gamma)\mathscr{B}_1(\gamma), C_2^{-1}(\gamma)\mathscr{B}_2(\gamma)] \\ \text{Cov}[C_2^{-1}(\gamma)\mathscr{B}_2(\gamma), C_1^{-1}(\gamma)\mathscr{B}_1(\gamma)] & \text{Var}[C_2^{-1}(\gamma)\mathscr{B}_2(\gamma)] \end{bmatrix}.$$

Start with $\text{Var}[\mathscr{B}_1(\gamma)]$. Write $v_t v_t^\top \otimes x_t x_t^\top$ as a short-cut for $(v_t v_t^\top) \otimes (x_t x_t^\top)$, and $\check{\theta}_z^{0\top} \otimes A^0 M_1(\gamma) M^{-1}$

as a short-cut for $\check{\theta}_z^{0\top} \otimes (A^0 M_1(\gamma) M^{-1})$. Then:

$$
\begin{aligned}
\mathrm{Var}[\mathscr{B}_1(\gamma)] &= \mathrm{Var}[A^0 \mathscr{G}\mathscr{P}_{\mathrm{mat},1}(\gamma)\tilde{\theta}_z^0 - A^0 M_1(\gamma) M^{-1} \mathscr{G}\mathscr{P}_{\mathrm{mat},1}\check{\theta}_z^0] \\
&= \mathrm{Var}[(\tilde{\theta}_z^{0\top} \otimes A^0)\mathscr{G}\mathscr{P}_1(\gamma)] + \mathrm{Var}[(\check{\theta}_z^{0\top} \otimes A^0 M_1(\gamma) M^{-1})\mathscr{G}\mathscr{P}] \\
&- \mathrm{Cov}[(\tilde{\theta}_z^{0\top} \otimes A^0)\mathscr{G}\mathscr{P}_1(\gamma), (\check{\theta}_z^{0\top} \otimes A^0 M_1(\gamma) M^{-1})\mathscr{G}\mathscr{P}] \\
&- \mathrm{Cov}[(\check{\theta}_z^{0\top} \otimes A^0 M_1(\gamma) M^{-1})\mathscr{G}\mathscr{P}, (\tilde{\theta}_z^{0\top} \otimes A^0)\mathscr{G}\mathscr{P}_1(\gamma)] \\
&= (\tilde{\theta}_z^{0\top} \otimes A^0)\mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t \le \gamma\}}](\tilde{\theta}_z^0 \otimes A^{0\top}) \\
&+ (\check{\theta}_z^{0\top} \otimes A^0 M_1(\gamma) M^{-1})\mathbb{E}[v_t v_t^\top \otimes x_t x_t^\top](\check{\theta}_z^0 \otimes M^{-1} M_1(\gamma) A^{0\top}) \\
&- (\tilde{\theta}_z^{0\top} \otimes A^0)\mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t \le \gamma\}}](\check{\theta}_z^0 \otimes M^{-1} M_1(\gamma) A^{0\top}) \\
&- (\check{\theta}_z^{0\top} \otimes A^0 M_1(\gamma) M^{-1})\mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t \le \gamma\}}](\tilde{\theta}_z^0 \otimes A^{0\top}) \\
&= A^0 \mathbb{E}[x_t x_t^\top (\epsilon_t + u_t^\top \theta_z^0)^2 \mathbb{1}_{\{q_t \le \gamma\}}] A^{0\top} \\
&+ A^0 M_1(\gamma) M^{-1} \mathbb{E}[x_t x_t^\top (u_t^\top \theta_z^0)^2] M^{-1} M_1(\gamma) A^{0\top} \\
&- A^0 \mathbb{E}[x_t x_t^\top (\epsilon_t u_t^\top \theta_z^0 + \theta_z^{0\top} u_t u_t^\top \theta_z^0)\mathbb{1}_{\{q_t \le \gamma\}}] M^{-1} M_1(\gamma) A^{0\top} \\
&- A^0 M_1(\gamma) M^{-1} \mathbb{E}[x_t x_t^\top](\epsilon_t u_t^\top \theta_z^0 + \theta_z^{0\top} u_t u_t^\top \theta_z^0)\mathbb{1}_{\{q_t \le \gamma\}}] A^{0\top},
\end{aligned}
$$

which yields the claim for $V_1(\gamma)$, when pre- and post-multiplied by $C_1^{-1}(\gamma)$.

Next, we consider $\mathrm{Var}[\mathscr{B}_2(\gamma)]$. First, note that

$$
\begin{aligned}
\mathscr{B}_2(\gamma) &= A^0 \mathscr{G}\mathscr{P}_{\mathrm{mat},1}\tilde{\theta}_z^0 - A^0 \mathscr{G}\mathscr{P}_{\mathrm{mat},1}\check{\theta}_z^0 - A^0 \mathscr{G}\mathscr{P}_{\mathrm{mat},1}(\gamma) + A^0 M_1(\gamma) M^{-1} \mathscr{G}\mathscr{P}_{\mathrm{mat},1}\check{\theta}_z^0 \\
&= A^0 \mathscr{G}\mathscr{P}_{\mathrm{mat},2}(\gamma)\tilde{\theta}_z^0 - A^0 M_2(\gamma) M^{-1} \mathscr{G}\mathscr{P}_{\mathrm{mat},1}\check{\theta}_z^0
\end{aligned}
$$

By similar arguments as for $\mathrm{Var}[\mathscr{B}_1(\gamma)]$,

$$
\begin{aligned}
\mathrm{Var}[\mathscr{B}_2(\gamma)] &= A^0 \mathbb{E}[x_t x_t^\top (\epsilon_t + u_t^\top \theta_z^0)^2 \mathbb{1}_{\{q_t > \gamma\}}] A^{0\top} \\
&+ A^0 M_2(\gamma) M^{-1} \mathbb{E}[x_t x_t^\top (u_t^\top \theta_z^0)^2] M^{-1} M_2(\gamma) A^{0\top} \\
&- A^0 \mathbb{E}[x_t x_t^\top (\epsilon_t u_t^\top \theta_z^0 + \theta_z^{0\top} u_t u_t^\top \theta_z^0)\mathbb{1}_{\{q_t > \gamma\}}] M^{-1} M_2(\gamma) A^{0\top} \\
&- A^0 M_2(\gamma) M^{-1} \mathbb{E}[x_t x_t^\top (\epsilon_t u_t^\top \theta_z^0 + \theta_z^{0\top} u_t u_t^\top \theta_z^0)\mathbb{1}_{\{q_t > \gamma\}}] A^{0\top}
\end{aligned}
$$

which yields the claim for $V_2(\gamma)$, when pre- and post-multiplied by $C_2^{-1}(\gamma)$.

Finally, we derive an expression for $\text{Cov}[\mathscr{B}_1(\gamma), \mathscr{B}_2(\gamma)]$[20]:

$$\begin{aligned}
\text{Cov}[\mathscr{B}_1(\gamma), \mathscr{B}_2(\gamma)] &= \text{Cov}[A^0 \mathscr{G}\mathscr{P}_{\text{mat},1}(\gamma)\tilde{\theta}_z^0 - A^0 M_1(\gamma) M^{-1} \mathscr{G}\mathscr{P}_{\text{mat},1}\check{\theta}_z^0, \\
&\qquad A^0 \mathscr{G}\mathscr{P}_{\text{mat},2}(\gamma)\tilde{\theta}_z^0 - A^0 M_2(\gamma) M^{-1} \mathscr{G}\mathscr{P}_{\text{mat},1}\check{\theta}_z^0] \\
&= -\text{Cov}[A^0 \mathscr{G}\mathscr{P}_{\text{mat},1}(\gamma)\tilde{\theta}_z^0, A^0 M_2(\gamma) M^{-1} \mathscr{G}\mathscr{P}_{\text{mat},1}\check{\theta}_z^0] \\
&\quad - \text{Cov}[A^0 M_1(\gamma) M^{-1} \mathscr{G}\mathscr{P}_{\text{mat},1}\check{\theta}_z^0, A^0 \mathscr{G}\mathscr{P}_{\text{mat},2}(\gamma)\tilde{\theta}_z^0] \\
&\quad + \text{Cov}[A^0 M_1(\gamma) M^{-1} \mathscr{G}\mathscr{P}_{\text{mat},1}\check{\theta}_z^0, A^0 M_2(\gamma) M^{-1} \mathscr{G}\mathscr{P}_{\text{mat},1}(\gamma)\check{\theta}_z^0] \\
&= -A^0 \mathbb{E}[x_t x_t^\top (\epsilon_t u_t^\top \theta_z^0 + \theta_z^{0\top} u_t u_t^\top \theta_z^0) \mathbb{1}_{\{q_t \leq \gamma\}}] M^{-1} M_2(\gamma) A^{0\top} \\
&\quad - A^0 M_1(\gamma) M^{-1} \mathbb{E}[x_t x_t^\top (\epsilon_t u_t^\top \theta_z^0 + \theta_z^{0\top} u_t u_t^\top \theta_z^0) \mathbb{1}_{\{q_t > \gamma\}}] A^{0\top} \\
&\quad + A M_1(\gamma) M^{-1} \mathbb{E}[x_t x_t^\top (u_t^\top \theta_z^0)^2] M^{-1} M_2(\gamma) A^{0\top}
\end{aligned}$$

which yields the claim for $V_{12}(\gamma)$ when pre-multiplied by $C_1^{-1}(\gamma)$ and post-multiplied by $C_2^{-1}(\gamma)$. □

Lemma 2.B.4. *Suppose Assumption 2.1 holds. Under $\mathbb{H}_0$ and uniformly in γ for $i = 1, 2$,*

$$\begin{aligned}
&(i)\ \hat{H}_i^{\epsilon}(\gamma) \xrightarrow{p} H_i^{\epsilon}(\gamma) && (ii)\ \hat{H}_i^{\epsilon,u}(\gamma) \xrightarrow{p} H_i^{\epsilon,u}(\gamma) \\
&(iii)\ \hat{H}_i^{u}(\gamma) \xrightarrow{p} H_i^{u}(\gamma) && (iv)\ \hat{H}_i(\gamma) \xrightarrow{p} H_i(\gamma)
\end{aligned}$$

PROOF OF LEMMA 2.B.4. Claim (i)**:** Note that, under $\mathbb{H}_0$, $\hat{\epsilon}_t = y_t - w_t^\top \hat{\theta}$ and start with

$$\begin{aligned}
\hat{H}_i^{\epsilon}(\gamma) &= T^{-1} \sum_{\mathscr{T}_i(\gamma)} x_t x_t^\top \hat{\epsilon}_t^2 \\
&= T^{-1} \sum_{\mathscr{T}_i(\gamma)} x_t x_t^\top (y_t - w_t^\top \hat{\theta})^2 \\
&= T^{-1} \sum_{\mathscr{T}_i(\gamma)} x_t x_t^\top [w_t^\top (\theta^0 - \hat{\theta}) + \epsilon_t]^2 \\
&= \underbrace{T^{-1} \sum_{\mathscr{T}_i(\gamma)} x_t x_t^\top [w_t^\top (\theta^0 - \hat{\theta})]^2}_{\text{(I)}} + \underbrace{2T^{-1} \sum_{\mathscr{T}_i(\gamma)} x_t x_t^\top \epsilon_t w_t^\top (\theta^0 - \hat{\theta})}_{\text{(II)}} + \underbrace{T^{-1} \sum_{\mathscr{T}_i(\gamma)} x_t x_t^\top \epsilon_t^2}_{\text{(III)}}.
\end{aligned}$$

We are left to show the limiting behavior of (I), (II), and (III).

[20] Note that $\text{Cov}[\mathscr{G}\mathscr{P}_1(\gamma), \mathscr{G}\mathscr{P}_2(\gamma)] = \mathbb{E}[\mathscr{G}\mathscr{P}_1(\gamma)\mathscr{G}\mathscr{P}_2^\top(\gamma)] = \mathbf{0}$.

$$
\begin{aligned}
\|(\mathrm{I})\|_F &\leq T^{-1} \sum_{\mathscr{T}_i(\gamma)} \|x_t x_t^\top [w_t^\top (\theta^0 - \hat{\theta})]^2\|_F \\
&\leq \left(T^{-1} \sum_{\mathscr{T}_i(\gamma)} \|x_t\|_2^2 \|w_t\|_2^2 \right) \|\theta^0 - \hat{\theta}\|_2^2 \\
&= \left(T^{-1} \sum_{\mathscr{T}_i(\gamma)} \|x_t\|_2^2 \|A^0 x_t + \bar{u}_t\|_2^2 \right) \|\theta^0 - \hat{\theta}\|_2^2 \\
&\leq \left(T^{-1} \sum_{T_i(\gamma)} \|x_t\|_2^2 \Big[\|A^0\|_F \|x_t\|_2 + \|u_t\|_2 \Big]^2 \right) \|\theta^0 - \hat{\theta}\|_2^2 \\
&= \left(T^{-1} \sum_{\mathscr{T}_i(\gamma)} \|x_t\|_2^4 \|A^0\|_F^2 + 2\|x_t\|_2^3 \|u_t\|_2 \|A^0\|_F + \|x_t\|_2^2 \|u_t\|_2^2 \right) \|\theta^0 - \hat{\theta}\|_2^2 \\
&= o_p(1) \qquad (2.39)
\end{aligned}
$$

where the last equality holds because $\|\theta^0 - \hat{\theta}\| = o_p(1)$ under $\mathbb{H}_0$ (follows directly from Lemma 2.B.2 by dropping γ) and the term in paranthesis is $\mathscr{O}_p(1)$. To see this latter claim, note that $\|A^0\|_F = \mathscr{O}_p(1)$ by Assumption 2.1 and consider

$$
\mathbb{P}\left(T^{-1} \sum_{\mathscr{T}_i(\gamma)} \|x_t\|_2^4 > K_1 \right) \leq \frac{\mathbb{E} \sum_{\mathscr{T}_i(\gamma)} \|x_t\|_2^4}{T K_1} \leq \frac{\sup_t \mathbb{E}\|x_t\|_2^4}{K_1}, \qquad (2.40a)
$$

$$
\begin{aligned}
\mathbb{P}\left(T^{-1} \sum_{\mathscr{T}_i(\gamma)} \|x_t\|_2^3 \|u_t\|_2 > K_2 \right) &\leq \frac{\mathbb{E} \sum_{\mathscr{T}_i(\gamma)} \|x_t\|_2^3 \|u_t\|_2}{T K_2} \leq \frac{\sup_t \mathbb{E}\|x_t\|_2^3 \|u_t\|_2}{K_2} \\
&\leq \frac{\sup_t \left[\mathbb{E}\|x_t\|_2^4 \right]^{3/4} \left[\mathbb{E}\|u_t\|_2^4 \right]^{1/4}}{K_2} \\
&\leq \frac{\sup_t \left[\mathbb{E}\|x_t\|_2^4 \right]^{3/4} \sup_t \left[\mathbb{E}\|u_t\|_2^4 \right]^{1/4}}{K_2} \qquad (2.40b)
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{P}\left(T^{-1} \sum_{\mathscr{T}_i(\gamma)} \|x_t\|_2^2 \|u_t\|_2^2 \right) &\leq \frac{\mathbb{E} \sum_{\mathscr{T}_i(\gamma)} \|x_t\|_2^2 \|u_t\|_2^2}{T K_3} \leq \frac{\sup_t \mathbb{E}\|x_t\|_2^2 \|u_t\|_2^2}{K_3} \\
&\leq \frac{\sup_t \left[\mathbb{E}\|x_t\|_2^4 \mathbb{E}\|u_t\|_2^4 \right]^{1/2}}{K_3} \leq \frac{\sup_t \left[\mathbb{E}\|x_t\|_2^4 \right]^{1/2} \sup_t \left[\mathbb{E}\|u_t\|_2^4 \right]^{1/2}}{K_3}. \qquad (2.40c)
\end{aligned}
$$

Now, by Assumption 2.1.2 it follows that all three terms (2.40a)–(2.40c) are $\mathscr{O}_p(1)$ and therefore, (2.39) follows.

For (II) it follows that

$$\begin{aligned}
\|(\text{II})\|_F &\leq T^{-1} \sum_{\mathcal{T}_i(\gamma)} \|x_t x_t^\top w_t^\top (\theta^0 - \hat{\theta})\epsilon_t\|_F \\
&\leq \Big(T^{-1} \sum_{\mathcal{T}_i(\gamma)} \|x_t\|_2^2 \|A^0 x_t + \bar{u}_t\|_2 |\epsilon_t| \Big) \|\theta^0 - \hat{\theta}\|_2 \\
&\leq \Big(T^{-1} \sum_{\mathcal{T}_i(\gamma)} \|x_t\|_2^3 \|A^0\|_F |\epsilon_t| + \|x_t\|_2^2 \|u_t\|_2 |\epsilon_t| \Big) \|\theta^0 - \hat{\theta}\|_2 \\
&= o_p(1). \qquad (2.41)
\end{aligned}$$

To see why this statement holds, consider

$$\begin{aligned}
\mathbb{P}\left(T^{-1} \sum_{\mathcal{T}_i(\gamma)} \|x_t\|_2^3 |\epsilon_t| > K_4 \right) &\leq \frac{\mathbb{E} \sum_{\mathcal{T}_i(\gamma)} \|x_t\|_2^3 |\epsilon_t|}{T K_4} \leq \frac{\sup_t \mathbb{E}\|x_t\|_2^3 |\epsilon_t|}{K_4} \\
&\leq \frac{\sup_t \left[\mathbb{E}\|x_t\|_2^4 \right]^{3/4} \left[\mathbb{E}|\epsilon_t|^4 \right]^{1/4}}{K_4} \\
&\leq \frac{\sup_t \left[\mathbb{E}\|x_t\|_2^4 \right]^{3/4} \sup_t \left[\mathbb{E}|\epsilon_t|^4 \right]^{1/4}}{K_4} \qquad (2.42a) \\
\mathbb{P}\left(T^{-1} \sum_{\mathcal{T}_i(\gamma)} \|x_t\|_2^2 \|u_t\|_2 |\epsilon_t| > K_5 \right) &\leq \frac{\mathbb{E} \sum_{\mathcal{T}_i(\gamma)} \|x_t\|_2^2 \|u_t\|_2 |\epsilon_t|}{T K_5} \leq \frac{\sup_t \mathbb{E}\|x_t\|_2^2 \|u_t \epsilon_t\|_2}{K_5} \\
&\leq \frac{\sup_t \left[\mathbb{E}\|x_t\|_2^4 \right]^{1/2} \left[\mathbb{E}\|u_t \epsilon_t\|_2^2 \right]^{1/2}}{K_5} \\
&\leq \frac{\sup_t \left[\mathbb{E}\|x_t\|_2^4 \right]^{1/2} \left[\mathbb{E}\|u_t\|_2^4 \right]^{1/4} \left[\mathbb{E}|\epsilon_t|^4 \right]^{1/4}}{K_5} \\
&\leq \frac{\sup_t \left[\mathbb{E}\|x_t\|_2^4 \right]^{1/2} \sup_t \left[\mathbb{E}\|u_t\|_2^4 \right]^{1/4} \sup_t \left[\mathbb{E}|\epsilon_t|^4 \right]^{1/4}}{K_5}. \qquad (2.42b)
\end{aligned}$$

Thus, the fact that $\|\theta^0 - \hat{\theta}\|_2 = o_p(1)$ together with (2.42a) and (2.42b) yield (2.41).
Finally, because (III) $\xrightarrow{p} H_1^c(\gamma)$, uniformly in γ, by Hansen (1996, Lemma 1), Claim (i) follows.

Claim (ii)**:** Under $\mathbb{H}_0$ and a LFS it holds that $\hat{\epsilon}_t = y_t - w_t^\top \hat{\theta}$ and $\hat{u}_t = z_t - \hat{\Pi}^\top x_t$. Therefore,

$$
\begin{aligned}
\hat{H}_i^{\epsilon,u}(\gamma) &= T^{-1} \sum_{\mathcal{T}_i(\gamma)} x_t x_t^\top \hat{\epsilon}_t \hat{u}_t^\top \hat{\theta}_z = T^{-1} \sum_{\mathcal{T}_i(\gamma)} x_t x_t^\top [w_t^\top (\theta^0 - \hat{\theta}) + \epsilon_t][x_t^\top (\Pi^0 - \hat{\Pi}) + u_t^\top]\hat{\theta}_z \\
&= \underbrace{T^{-1} \sum_{\mathcal{T}_i(\gamma)} x_t x_t^\top x_t^\top A^0 (\theta^0 - \hat{\theta}) x_t^\top (\Pi^0 - \hat{\Pi})\hat{\theta}_z}_{\text{(IV)}} + \underbrace{T^{-1} \sum_{\mathcal{T}_i(\gamma)} x_t x_t^\top x_t^\top A^0 (\theta^0 - \hat{\theta}) u_t^\top \hat{\theta}_z}_{\text{(V)}} \\
&+ \underbrace{T^{-1} \sum_{\mathcal{T}_i(\gamma)} x_t x_t^\top \bar{u}_t^\top (\theta^0 - \hat{\theta}) x_t^\top (\Pi^0 - \hat{\Pi})\hat{\theta}_z}_{\text{(VI)}} + \underbrace{T^{-1} \sum_{\mathcal{T}_i(\gamma)} x_t x_t^\top \bar{u}_t^\top (\theta^0 - \hat{\theta}) u_t^\top \hat{\theta}_z}_{\text{(VII)}} \\
&+ \underbrace{T^{-1} \sum_{\mathcal{T}_i(\gamma)} x_t x_t^\top \epsilon_t x_t^\top (\Pi^0 - \hat{\Pi})\hat{\theta}_z}_{\text{(VIII)}} + \underbrace{T^{-1} \sum_{\mathcal{T}_i(\gamma)} x_t x_t^\top \epsilon_t^\top u_t^\top \hat{\theta}_z}_{\text{(IX)}}. \qquad (2.43)
\end{aligned}
$$

Now, it immediately follows, by similar arguments as for Claim (i), that

$$
\begin{aligned}
\|\text{(IV)}\|_F &\leq \Big(T^{-1} \sum_{\mathcal{T}_i(\gamma)} \|x_t\|_2^4\Big) \|A^0\|_F \|\Pi^0 - \hat{\Pi}\|_F \|\theta^0 - \hat{\theta}\|_2 \|\hat{\theta}_z\|_2 \\
&= \mathcal{O}_p(1)\mathcal{O}_p(1)o_p(1)o_p(1)\mathcal{O}_p(1) = o_p(1) \qquad (2.44a) \\
\|\text{(V)}\|_F &\leq \Big(T^{-1} \sum_{\mathcal{T}_i(\gamma)} \|x_t\|_2^3 \|u_t\|_2\Big) \|A^0\|_F \|\theta^0 - \hat{\theta}\|_2 \|\hat{\theta}_z\|_2 \\
&= \mathcal{O}_p(1)\mathcal{O}_p(1)o_p(1)\mathcal{O}_p(1) = o_p(1) \qquad (2.44b) \\
\|\text{(VI)}\|_F &\leq \Big(T^{-1} \sum_{\mathcal{T}_i(\gamma)} \|x_t\|_2^3 \|u_t\|_2\Big) \|\Pi^0 - \hat{\Pi}\|_F \|\theta^0 - \hat{\theta}\|_2 \|\hat{\theta}_z\|_2 \\
&= \mathcal{O}_p(1)o_p(1)o_p(1)\mathcal{O}_p(1) = o_p(1) \qquad (2.44c) \\
\|\text{(VII)}\|_F &\leq \Big(T^{-1} \sum_{\mathcal{T}_i(\gamma)} \|x_t\|_2^2 \|u_t\|_2^2\Big) \|\theta^0 - \hat{\theta}\|_2 \|\hat{\theta}_z\|_2 \\
&= \mathcal{O}_p(1)o_p(1)\mathcal{O}_p(1) = o_p(1) \qquad (2.44d) \\
\|\text{(VIII)}\|_F &\leq \Big(T^{-1} \sum_{\mathcal{T}_i(\gamma)} \|x_t\|_2^3 |\epsilon_t|\Big) \|\Pi^0 - \hat{\Pi}\|_F \|\hat{\theta}_z\|_2 \\
&= \mathcal{O}_p(1)o_p(1)\mathcal{O}_p(1) = o_p(1). \qquad (2.44e)
\end{aligned}
$$

For the last term in (2.43) it holds, uniformly in γ by Hansen (1996, Lemma 1), that (IX) $= T^{-1} \sum_{\mathcal{T}_i(\gamma)} x_t x_t^\top \epsilon_t u_t^\top (\theta^0 + o_p(1)) = T^{-1} \sum_{\mathcal{T}_i(\gamma)} x_t x_t^\top \epsilon_t u_t^\top \theta^0 + o_p(1) \xrightarrow{p} H_1^{\epsilon,u}(\gamma)$. Thus, Claim (ii) follows together with (2.44a)–(2.44e).

Claim (iii)**:** As before, $\hat{u}_t = z_t - \hat{\Pi}^\top x_t$. Then, under $\mathbb{H}_0$, it follows that

$$\begin{aligned}\hat{H}_i^u(\gamma) &= T^{-1}\sum_{\mathcal{T}_i(\gamma)} x_t x_t^\top (\hat{u}_t^\top \hat{\theta}_z)^2 = T^{-1}\sum_{\mathcal{T}_i(\gamma)} x_t x_t^\top \{[x_t^\top(\Pi^0-\hat{\Pi}) + u_t^\top]\hat{\theta}_z\}^2 \\ &= \underbrace{T^{-1}\sum_{\mathcal{T}_i(\gamma)} x_t x_t^\top [x_t^\top(\Pi^0-\hat{\Pi})\hat{\theta}_z]^2}_{\text{(X)}} + \underbrace{2T^{-1}\sum_{\mathcal{T}_i(\gamma)} x_t x_t^\top x_t^\top(\Pi^0-\hat{\Pi})\hat{\theta}_z u_t^\top \hat{\theta}_z}_{\text{(XI)}} \\ &\quad + \underbrace{T^{-1}\sum_{\mathcal{T}_i(\gamma)} x_t x_t^\top (u_t^\top \hat{\theta}_z)^2}_{\text{(XII)}} \end{aligned} \tag{2.45}$$

Next,

$$\|\text{(X)}\|_F \leq \Big(T^{-1}\sum_{\mathcal{T}_i(\gamma)} \|x_t\|_2^4\Big)\|\Pi^0-\hat{\Pi}\|_F^2\|\hat{\theta}_z\|_2^2 = \mathcal{O}_p(1)o_p(1)\mathcal{O}_p(1) = o_p(1) \tag{2.46a}$$

$$\|\text{(XI)}\|_F \leq \Big(T^{-1}\sum_{\mathcal{T}_i(\gamma)} \|x_t\|_2^3\|u_t\|_2\Big)\|\Pi^0-\hat{\Pi}\|_F\|\hat{\theta}_z\|_2^2 = \mathcal{O}_p(1)o_p(1)\mathcal{O}_p(1) = o_p(1). \tag{2.46b}$$

For the last term in (2.45) it holds, uniformly in γ by Hansen (1996, Lemma 1), that (X) $= T^{-1}\sum_{\mathcal{T}_i(\gamma)} x_t x_t^\top (u_t^\top \hat{\theta}_z)^2 = T^{-1}\sum_{\mathcal{T}_i(\gamma)} x_t x_t^\top (u_t^\top \theta_z^0)^2 + o_p(1) \xrightarrow{p} H^u(\gamma)$. Thus, Claim (iii) follows together with (2.46a) and (2.46b).

Claim (iv)**:** This claim follows by noting that $\hat{H}_i(\gamma) = \hat{H}_i^\epsilon(\gamma) + 2\hat{H}_i^{\epsilon,u}(\gamma) + \hat{H}_i^u(\gamma)$, using Claims (i)–(iii) and the continuous mapping theorem. □

PROOF OF THEOREM 2.2.

(i) sup LR Test: This proof is done in two parts: part (A) shows that $T^{-1}SSR_1(\gamma) \xrightarrow{p} \sigma^2$ and part (B) shows that $SSR_0 - SSR_1(\gamma) \Rightarrow \mathscr{E}^\top(\gamma)C_2(\gamma)C^{-1}C_1(\gamma)\mathscr{E}(\gamma)$.

Part (A)**.** The scaled sum of squared residuals of the restricted model, $SSR_1(\gamma)$, is

$$\begin{aligned} T^{-1}SSR_1(\gamma) &= T^{-1}[Y_1^\gamma - \hat{W}_1^\gamma\hat{\theta}_1^\gamma]^\top[Y_1^\gamma - \hat{W}_1^\gamma\hat{\theta}_1^\gamma] \\ &\quad + T^{-1}[Y_2^\gamma - \hat{W}_2^\gamma\hat{\theta}_2^\gamma]^\top[Y_2^\gamma - \hat{W}_2^\gamma\hat{\theta}_2^\gamma] \\ &= T^{-1}[\hat{W}_1^\gamma(\theta^0-\hat{\theta}_1^\gamma) + \tilde{\epsilon}_1^\gamma]^\top[\hat{W}_1^\gamma(\theta^0-\hat{\theta}_1^\gamma) + \tilde{\epsilon}_1^\gamma] \\ &\quad + T^{-1}[\hat{W}_2^\gamma(\theta^0-\hat{\theta}_2^\gamma) + \tilde{\epsilon}_2^\gamma]^\top[\hat{W}_2^\gamma(\theta^0-\hat{\theta}_2^\gamma) + \tilde{\epsilon}_2^\gamma] \\ &= T^{-1}\tilde{\epsilon}^\top\tilde{\epsilon} \\ &\quad + 2(T^{-1}\tilde{\epsilon}_1^{\gamma\top}\hat{W}_1^\gamma)(\theta^0-\hat{\theta}_1^\gamma) + (\theta^0-\hat{\theta}_1^\gamma)^\top(T^{-1}\hat{W}_1^{\gamma\top}\hat{W}_1^\gamma)(\theta^0-\hat{\theta}_1^\gamma) \\ &\quad + 2(T^{-1}\tilde{\epsilon}_2^{\gamma\top}\hat{W}_2^\gamma)(\theta^0-\hat{\theta}_2^\gamma) + (\theta^0-\hat{\theta}_2^\gamma)^\top(T^{-1}\hat{W}_2^{\gamma\top}\hat{W}_2^\gamma)(\theta^0-\hat{\theta}_2^\gamma). \end{aligned} \tag{2.47}$$

Next, by Lemma 2.B.2, for $i = 1,2$, $T^{-1}\hat{W}_i^{\gamma\top}\tilde{\epsilon}_i^\gamma = o_p(1)$ and $T^{-1}\hat{W}_i^{\gamma\top}\hat{W}_i^\gamma = \mathcal{O}_p(1)$ uniformly in γ. This implies that

$$\hat{\theta}_i^\gamma - \theta^0 = (T^{-1}\hat{W}_i^{\gamma\top}\hat{W}_i^\gamma)^{-1}(\hat{W}_i^{\gamma\top}\tilde{\epsilon}_i^\gamma) = \mathcal{O}_p(1)o_p(1) = o_p(1)$$

and therefore, (2.47) simplifies to

$$\begin{aligned} T^{-1}SSR_1(\gamma) &= T^{-1}\tilde{\epsilon}^\top\tilde{\epsilon} + o_p(1) \\ &= T^{-1}s^\top s + 2(T^{-1}s^\top X)(\Pi^0 - \hat{\Pi})\theta_z^0 \\ &\quad + \theta_z^{0\top}(\Pi^0 - \hat{\Pi})X^\top X(\Pi^0 - \hat{\Pi})\theta_z^0 + o_p(1) \end{aligned} \tag{2.48}$$

where $s = \epsilon + u^\top\theta_z^0$, $\hat{\Pi} - \Pi^0 = o_p(1)$ and $T^{-1}s^\top X = o_p(1)$ by Lemma 2.B.2, uniformly in γ. Thus, (2.48) simplifies to

$$\begin{aligned} T^{-1}SSR_1(\gamma) &= T^{-1}s^\top s + o_p(1) \\ &= T^{-1}\epsilon^\top\epsilon + 2(T^{-1}\epsilon^\top u)\theta_z^0 + \theta_z^{0\top}(T^{-1}u^\top u)\theta_z^0 + o_p(1) \\ &\xrightarrow{p} \sigma_\epsilon^2 + 2\Sigma_{\epsilon,u}^\top\theta_z^0 + \theta_z^{0\top}\Sigma_u\theta_z^0 = \sigma^2 \end{aligned}$$

uniformly in γ. This proves part (i).

Part (B)**.** We have that

$$\begin{aligned} SSR_0 - SSR_1(\gamma) &= [Y_1^\gamma - \hat{W}_1^\gamma\hat{\theta}]^\top[Y_1^\gamma - \hat{W}_1^\gamma\hat{\theta}] - [Y_1^\gamma - \hat{W}_1^\gamma\hat{\theta}_1^\gamma]^\top[Y_1^\gamma - \hat{W}_1^\gamma\hat{\theta}_1^\gamma] \\ &\quad + [Y_2^\gamma - \hat{W}_2^\gamma\hat{\theta}]^\top[Y_2^\gamma - \hat{W}_2^\gamma\hat{\theta}] - [Y_2^\gamma - \hat{W}_2^\gamma\hat{\theta}_2^\gamma]^\top[Y_2^\gamma - \hat{W}_2^\gamma\hat{\theta}_2^\gamma] \end{aligned} \tag{2.49}$$

Now, for $i = 1,2$,

$$\begin{aligned} &[Y_i^\gamma - \hat{W}_i^\gamma\hat{\theta}]^\top[Y_i^\gamma - \hat{W}_i^\gamma\hat{\theta}] \\ -[Y_i^\gamma - \hat{W}_i^\gamma\hat{\theta}_i^\gamma]^\top[Y_i^\gamma - \hat{W}_i^\gamma\hat{\theta}_i^\gamma] &= Y_i^{\gamma\top}Y_i^\gamma - 2\hat{\theta}^\top\hat{W}_i^{\gamma\top}Y_i^\gamma + \hat{\theta}^\top\hat{W}_i^{\gamma\top}\hat{W}_i^\gamma\hat{\theta} \\ &\quad - Y_i^{\gamma\top}Y_i^\gamma + 2\hat{\theta}_i^{\gamma\top}\hat{W}_i^\gamma - \hat{\theta}_i^{\gamma\top}\hat{W}_i^{\gamma\top}\hat{W}_i^\gamma\hat{\theta}_i^\gamma \\ &= [\hat{\theta}_i^\gamma - \hat{\theta}]^\top\hat{W}_i^{\gamma\top}[2Y_i^\gamma - \hat{W}_i^\gamma\hat{\theta} - \hat{W}_i^\gamma\hat{\theta}_i^\gamma] \\ &= T^{1/2}[\hat{\theta}_i^\gamma - \hat{\theta}]^\top\Big[2(T^{-1/2}\hat{W}_i^{\gamma\top}\tilde{\epsilon}_i^\gamma) \\ &\quad -(T^{-1}\hat{W}_i^{\gamma\top}\hat{W}_i^\gamma)(T^{1/2}(\hat{\theta} - \theta^0)) \\ &\quad -(T^{-1}\hat{W}_i^{\gamma\top}\hat{W}_i^\gamma)(T^{1/2}(\hat{\theta}_i^\gamma - \theta^0))\Big]. \end{aligned} \tag{2.50}$$

Next, we show the asymptotic behavior of the terms on the right hand side of (2.50) which then concludes the proof together with Part (i), (2.49), the continuous mapping theorem and weak convergence (uniformly in γ). It holds that

$$\begin{aligned} &(T^{-1}\hat{W}^\top\hat{W})(T^{1/2}(\hat{\theta} - \theta^0)) \\ =&T^{-1/2}\hat{W}^\top\tilde{\epsilon} \\ =&T^{-1/2}\hat{W}_1^{\gamma\top}\tilde{\epsilon}_1^\gamma + T^{-1/2}\hat{W}_2^{\gamma\top}\tilde{\epsilon}_2^\gamma \\ =&(T^{-1}\hat{W}_1^{\gamma\top}\hat{W}_1^\gamma)(T^{1/2}(\hat{\theta}_1^\gamma - \theta^0)) + (T^{-1}\hat{W}_2^{\gamma\top}\hat{W}_2^\gamma)(T^{1/2}(\hat{\theta}_2^\gamma - \theta^0)) \end{aligned} \tag{2.51}$$

and by Lemma 2.B.2 that, uniformly in γ for $i = 1,2$,

$$T^{-1}\hat{W}_i^{\gamma\top}\hat{W}_i^\gamma \xrightarrow{p} C_i(\gamma). \tag{2.52}$$

Define $\hat{\beta} \equiv T^{1/2}(\hat{\theta} - \theta^0)$, $\hat{\beta}_i \equiv T^{1/2}(\hat{\theta}_i^\gamma - \theta^0)$ and $D_i(\gamma) \equiv C^{-1}C_i(\gamma)$ $(i = 1,2)$. Then, (2.52) can be restated as

$$\hat{\beta} = D_1(\gamma)\hat{\beta}_1 + D_2(\gamma)\hat{\beta}_2 + o_p(1). \tag{2.53}$$

Moreover, note that because $D_1(\gamma) + D_2(\gamma) = I$,

$$T^{1/2}(\hat{\theta}_1^\gamma - \hat{\theta}) = \hat{\beta}_1 - \hat{\beta} = D_2(\gamma)(\hat{\beta}_1 - \hat{\beta}_2) + o_p(1) \tag{2.54a}$$

$$T^{1/2}(\hat{\theta}_2^\gamma - \hat{\theta}) = \hat{\beta}_2 - \hat{\beta} = -D_1(\gamma)(\hat{\beta}_1 - \hat{\beta}_2) + o_p(1) \tag{2.54b}$$

$$T^{-1/2}\hat{W}_i^{\gamma\top}\tilde{\epsilon}_i^\gamma = C_i(\gamma)\hat{\beta}_i + o_p(1) \tag{2.54c}$$

by (2.53) and Lemma 2.B.2.
So, using (2.51)–(2.54c), quantity (2.50) can be written, for $i = 1$, as

$$\begin{aligned}
&(\hat{\beta}_1 - \hat{\beta}_2)^\top D_2^\top(\gamma)\left[2C_1(\gamma)\hat{\beta}_1 - C_1(\gamma)\hat{\beta} - C_1(\gamma)\hat{\beta}_1\right] + o_p(1) \\
=&(\hat{\beta}_1 - \hat{\beta}_2)^\top D_2^\top(\gamma)C_1(\gamma)(\hat{\beta}_1 - \hat{\beta}) + o_p(1) \\
=&(\hat{\beta}_1 - \hat{\beta}_2)^\top D_2^\top(\gamma)C_1(\gamma)D_2(\gamma)(\hat{\beta}_1 - \hat{\beta}_2) + o_p(1). \qquad (2.55)
\end{aligned}$$

Similarly, using (2.51)–(2.53) and (2.54b), and(2.54c), quantity (2.50) can be stated, for $i = 2$, as

$$(\hat{\beta}_1 - \hat{\beta}_2)^\top D_1^\top(\gamma)C_2(\gamma)D_1(\gamma)(\hat{\beta}_1 - \hat{\beta}_2) + o_p(1). \tag{2.56}$$

So, using (2.50), (2.55) and (2.56), quantity (2.49) can be restated as

$$\begin{aligned}
SSR_0 - SSR_1(\gamma) &= (\hat{\beta}_1 - \hat{\beta}_2)^\top D_2^\top(\gamma)C_1(\gamma)D_2(\gamma)(\hat{\beta}_1 - \hat{\beta}_2) \\
&\quad + (\hat{\beta}_1 - \hat{\beta}_2)^\top D_1^\top(\gamma)C_2(\gamma)D_1(\gamma)(\hat{\beta}_1 - \hat{\beta}_2) + o_p(1) \\
&= (\hat{\beta}_1 - \hat{\beta}_2)^\top \left[(I_p - D_1^\top(\gamma))C_1(\gamma)(I_p - D_1(\gamma))\right. \\
&\quad \left. + D_1^\top(\gamma)(C - C_1(\gamma))D_1(\gamma)\right](\hat{\beta}_1 - \hat{\beta}_2) + o_p(1) \\
&= (\hat{\beta}_1 - \hat{\beta}_2)^\top \left[C_1(\gamma) - 2C_1(\gamma)D_1(\gamma) + D_1^\top(\gamma)C_1(\gamma)D_1(\gamma)\right. \\
&\quad \left. + D_1^\top(\gamma)CD_1(\gamma) - D_1^\top(\gamma)C_1(\gamma)D_1(\gamma)\right](\hat{\beta}_1 - \hat{\beta}_2) + o_p(1) \\
&= (\hat{\beta}_1 - \hat{\beta}_2)^\top \left[C_1(\gamma) - C_1(\gamma)D_1(\gamma)\right](\hat{\beta}_1 - \hat{\beta}_2) + o_p(1) \\
&= (\hat{\beta}_1 - \hat{\beta}_2)^\top C_2(\gamma)D_1(\gamma)(\hat{\beta}_1 - \hat{\beta}_2) + o_p(1). \qquad (2.57)
\end{aligned}$$

Last, by Lemma 2.B.2 it holds, uniformly in γ, that

$$\begin{aligned}
\hat{\beta}_1 - \hat{\beta}_2 &= (T^{-1}\hat{W}_1^{\gamma\top}\hat{W}_1^\gamma)^{-1}(T^{-1/2}\hat{W}_1^{\gamma\top}\tilde{\epsilon}_1^\gamma) - (T^{-1}\hat{W}_2^{\gamma\top}\hat{W}_2^\gamma)^{-1}(T^{-1/2}\hat{W}_2^{\gamma\top}\tilde{\epsilon}_2^\gamma) \\
&\Rightarrow C_1^{-1}(\gamma)A^0\mathscr{B}_1(\gamma) - C_2^{-1}(\gamma)A^0\mathscr{B}_2(\gamma) \equiv \mathscr{E}(\gamma). \qquad (2.58)
\end{aligned}$$

So, combining (2.57) and (2.58) yields

$$SSR_0 - SSR_1(\gamma) \Rightarrow \mathscr{E}^\top(\gamma)C_2(\gamma)D_1(\gamma)\mathscr{E}(\gamma)$$

which in turn with Part (A), the continuous mapping theorem and weak convergence (uniformly in γ) proves the claim.

(ii) sup Wald Test: From Equation (2.50) it follows that

$$T^{1/2}(\hat{\theta}_1^\gamma - \hat{\theta}_2^\gamma) \Rightarrow \mathscr{E}(\gamma). \tag{2.59}$$

Moreover, from Definition 2.2,

$$\begin{aligned}\hat{V}_i(\gamma) = \hat{C}_i^{-1}(\gamma)\hat{A}\Big[\hat{H}_i(\gamma) + \hat{R}_i(\gamma)\hat{H}^u\hat{R}_i^\top(\gamma) - [\hat{H}_i^{\epsilon,u}(\gamma) + \hat{H}_i^u(\gamma)]\hat{R}_i^\top(\gamma) \\ - \hat{R}_i(\gamma)[\hat{H}_i^{\epsilon,u}(\gamma) + \hat{H}_i^u(\gamma)]\Big]\hat{A}^\top\hat{C}_i^{-1}(\gamma)\end{aligned} \tag{2.60a}$$

and

$$\begin{aligned}\hat{V}_{12}(\gamma) = \hat{C}_1^{-1}(\gamma)\hat{A}\Big[(\hat{H}_1^{\epsilon,u}(\gamma) + \hat{H}_1^u(\gamma))\hat{R}_2^\top(\gamma) - \hat{R}_1(\gamma)(\hat{H}_2^{\epsilon,u}(\gamma) + \hat{H}_2^u(\gamma)) \\ + \hat{R}_1(\gamma)\hat{H}^u\hat{R}_2^\top(\gamma))\Big]\hat{A}^\top\hat{C}_2^{-1}(\gamma)\end{aligned} \tag{2.60b}$$

Now, by (2.29) and the continuous mapping theorem it immediately follows, uniformly in γ, that

$$\hat{R}_i(\gamma) = \hat{M}_i(\gamma)\hat{M}^{-1} = \Big(T^{-1}X_i^{\gamma\top}X_i^\gamma\Big)\Big(T^{-1}X^\top X\Big)^{-1} \xrightarrow{p} M_i(\gamma)M^{-1} = R_i(\gamma). \tag{2.61}$$

Moreover, by Lemma 2.B.2 and the continuous mapping theorem it also holds, uniformly in γ, that

$$\hat{C}_i^{-1}(\gamma) = (T^{-1}\hat{W}_i^{\gamma\top}\hat{W}_i^\gamma)^{-1} = \Big(\hat{A}\hat{M}_i(\gamma)\hat{A}^\top\Big)^{-1} \xrightarrow{p} C_i^{-1}(\gamma), \quad \text{and} \quad \hat{A} = [\hat{\Pi}, S^\top]^\top \xrightarrow{p} A^0. \tag{2.62}$$

Finally, in Lemma 2.B.4 we derived the limits of $\hat{H}_i^\epsilon(\gamma)$, $\hat{H}_i^u(\gamma)$ and $\hat{H}_i^{\epsilon,u}(\gamma)$ concluding the proof together with (2.59)–(2.62). □

Corollary 2.B.1 (to Theorem 2.2)**.** *Let Z be generated by* (2.1)*, Y be generated by* (2.3)*, and $\hat{Z}$ be calculated by* (2.4)*. Under $\mathbb{H}_0$ and Assumptions 2.1-2.2,*

(*i*)

$$\sup_{\gamma\in\Gamma} LR_{T,LFS}^{2SLS}(\gamma) \Rightarrow \sup_{\gamma\in\Gamma} \tilde{\mathscr{E}}^\top(\gamma)\mathcal{Q}^{-1}(\gamma)\tilde{\mathscr{E}}(\gamma)$$

(*ii*)

$$\sup_{\gamma\in\Gamma} W_{T,LFS}^{2SLS}(\gamma) \Rightarrow \sup_{\gamma\in\Gamma} \tilde{\mathscr{E}}^\top(\gamma)\tilde{V}^{-1}(\gamma)\tilde{\mathscr{E}}(\gamma)$$

where $\tilde{V}(\gamma) = \tilde{V}_1(\gamma) + \tilde{V}_2(\gamma) - \tilde{V}_{12}(\gamma) - \tilde{V}_{12}^\top(\gamma)$,

$$\begin{aligned}\tilde{V}_i(\gamma) &= C_i^{-1}(\gamma)\, A_0\Big[\sigma^2 I_q - (\sigma^2 - \sigma_\epsilon^2)R_i(\gamma)\Big]M_i(\gamma)A_0^\top\, C_i^{-1}(\gamma) \\ \tilde{V}_{12}(\gamma) &= -C_1^{-1}(\gamma)(\sigma^2 - \sigma_\epsilon^2)A^0R_1(\gamma)M_2(\gamma)A^{0\top}C_2^{-1}(\gamma),\end{aligned}$$

and $\widetilde{\mathscr{GP}}_{\text{mat},1}(\gamma)$ is a $q \times (p_1+1)$-matrix where all columns are independent $q \times 1$ zero mean Gaussian processes with covariance kernel[21] *$M_1(\gamma_1 \wedge \gamma_2)$, $\Sigma^{1/2}$ is the principal square root of*

[21] Thus, the only difference between the two Gaussian processes $\widetilde{\mathscr{GP}}_{\text{mat},1}(\gamma)$ and $\mathscr{GP}_{\text{mat},1}(\gamma)$ lies in their covariance functions.

Σ, $\tilde{\mathscr{E}}(\gamma) = C_1^{-1}(\gamma)\tilde{\mathscr{B}}_1(\gamma) - C_2^{-1}(\gamma)\tilde{\mathscr{B}}_2(\gamma)$, *and* $\tilde{\mathscr{B}}_1(\gamma) = A^0[\tilde{\mathscr{GP}}_{\text{mat},1}(\gamma)\Sigma^{1/2}\tilde{\theta}_z^0 - R_1(\gamma)\tilde{\mathscr{GP}}_{\text{mat}}\Sigma^{1/2}\check{\theta}_z^0]$, $\tilde{\mathscr{B}}_2(\gamma) = \tilde{\mathscr{B}}_1(\gamma_{max}) - \tilde{\mathscr{B}}_1(\gamma)$.

(iii) If the system is just-identified, i.e. if $p = q$, then the two test statistics are asymptotically equivalent with asymptotic distribution given by $\sup_{\gamma\in\Gamma} J_1(\gamma)$*, where:*

$$\begin{aligned} J_1(\gamma) = \frac{1}{\sigma^2}(\Sigma^{1/2}\tilde{\theta}_z^0)^\top [M_1^{-1}(\gamma)\tilde{\mathscr{GP}}_{\text{mat},1}(\gamma) - M_2^{-1}(\gamma)\tilde{\mathscr{GP}}_{\text{mat},2}(\gamma)]^\top \\ \times [M_2(\gamma)M^{-1}M_1(\gamma)] \\ \times [M_1^{-1}(\gamma)\tilde{\mathscr{GP}}_{\text{mat},1}(\gamma) - M_2^{-1}(\gamma)\tilde{\mathscr{GP}}_{\text{mat},2}(\gamma)]\Sigma^{1/2}\tilde{\theta}_z^0. \end{aligned}$$

PROOF OF COROLLARY 2.B.1.

(i) sup LR-test: We only need to show $\mathscr{E}(\gamma) = \tilde{\mathscr{E}}(\gamma)$ under Assumptions 2.1 and 2.2; in other words, that $\mathscr{GP}_{\text{mat},1}(\gamma) = \tilde{\mathscr{GP}}_{\text{mat},1}(\gamma)\Sigma^{1/2}$.[22] The covariance kernel of $\mathscr{GP}_1(\gamma)$ is given as $\mathbb{E}[\mathscr{GP}_1(\gamma_1)\mathscr{GP}_1^\top(\gamma_2)] = \mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t\le\gamma_1\wedge\gamma_2\}}]$ by Lemma 2.B.1, using the shortcut notation $v_t v_t^\top \otimes x_t x_t^\top = (v_t v_t^\top)\otimes(x_t x_t^\top)$. Under Assumption 2.2 this expression can be simplified to

$$\begin{aligned} \mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t\le\gamma_1\wedge\gamma_2\}}] &= \mathbb{E}\left[\mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t\le(\gamma_1\wedge\gamma_2)\}}|x_t, q_t]\right] \\ &= \mathbb{E}\left[\mathbb{E}[v_t v_t^\top|x_t, q_t]\otimes(x_t x_t^\top\mathbb{1}_{\{q_t\le(\gamma_1\wedge\gamma_2)\}})\right] \\ &= \mathbb{E}\left[\Sigma\otimes(x_t x_t^\top\mathbb{1}_{\{q_t\le(\gamma_1\wedge\gamma_2)\}})\right] \\ &= \Sigma\otimes M_1(\gamma_1\wedge\gamma_2). \end{aligned}$$

Next, the principal square root of Σ, i.e. $\Sigma^{1/2}$ that satisfies $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$, exists since Σ is positive definite by Assumption 2.1.5. Thus,

$$\begin{aligned} \mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t\le\gamma_1\wedge\gamma_2\}}] &= \Sigma\otimes M_1(\gamma_1\wedge\gamma_2) \\ &= (\Sigma^{1/2}\otimes M_1(\gamma_1\wedge\gamma_2))(\Sigma^{1/2}\otimes I_q) \\ &= (\Sigma^{1/2}\otimes I_q)(I_{p_1+1}\otimes M_1(\gamma_1\wedge\gamma_2))(\Sigma^{1/2}\otimes I_q). \end{aligned} \tag{2.63}$$

The covariance kernel of $(\Sigma^{1/2}\otimes I_q)\tilde{\mathscr{GP}}_1(\gamma) = \text{vec}(\tilde{\mathscr{GP}}_{\text{mat},1}(\gamma)\Sigma^{1/2})$ is given by

$$\begin{aligned} \mathbb{E}[(\Sigma^{1/2}\otimes I_q)\tilde{\mathscr{GP}}_1(\gamma_1)\tilde{\mathscr{GP}}_1^\top(\gamma_2)(\Sigma^{1/2}\otimes I_q)] &= (\Sigma^{1/2}\otimes I_q)\mathbb{E}[\tilde{\mathscr{GP}}_1(\gamma_1)\tilde{\mathscr{GP}}_1^\top(\gamma_2)](\Sigma^{1/2}\otimes I_q) \\ &= (\Sigma^{1/2}\otimes I_q)(I_{p_1+1}\otimes M_1(\gamma_1\wedge\gamma_2))(\Sigma^{1/2}\otimes I_q) \end{aligned} \tag{2.64}$$

because $\mathbb{E}[\tilde{\mathscr{GP}}_1(\gamma_1)\tilde{\mathscr{GP}}_1^\top(\gamma_2)] = I_{p_1+1}\otimes M_1(\gamma_1\wedge\gamma_2)$ by definition of $\tilde{\mathscr{GP}}_1(\gamma)$. Combining (2.63) and (2.64) yields the desired result since Gaussian processes are uniquely defined through their mean and covariance functions.

[22] We will do this by showing that their covariance functions are the same. Hence, because both processes have mean zero, equality follows due to the fact that Gaussian processes are uniquely defined through their mean and covariance functions.

(ii) sup Wald-test: Conditional homoskedasticity implies that $H_i^{\epsilon}(\gamma) = \sigma_{\epsilon}^2 M_i(\gamma)$, $H_i^{u}(\gamma) = (\theta_z^{0\top}\Sigma_u\theta_z^0)\, M_i(\gamma)$, and $H_i^{\epsilon,u}(\gamma) = (\Sigma_{\epsilon,u}^{\top}\theta_z^0)\, M_i(\gamma)$. Plugging these results into the expression for $V(\gamma)$ in Definition 2.2 and simplifying yields the asymptotic distribution of the sup-Wald test for the overidentified case.

(iii) To show the asymptotic equivalence for $p = q$, define $\Delta_\sigma = \sigma^2 - \sigma_{\epsilon}^2$. Then:

$$\begin{aligned}
\tilde{V}(\gamma) &= \tilde{V}_1(\gamma) + \tilde{V}_2(\gamma) - \tilde{V}_{12}(\gamma) - \tilde{V}_{12}^{\top}(\gamma) \\
&= \sigma^2 C_1^{-1}(\gamma) - \Delta_\sigma C_1^{-1}(\gamma) A^0 R_1(\gamma) M_1(\gamma) A^{0\top} C_1^{-1}(\gamma) \\
&\quad + \sigma^2 C_2^{-1}(\gamma) - \Delta_\sigma C_2^{-1}(\gamma) A^0 R_2(\gamma) M_2(\gamma) A^{0\top} C_2^{-1}(\gamma) \\
&\quad + \Delta_\sigma C_1^{-1}(\gamma) A^0 R_1(\gamma) M_2(\gamma) A^{0\top} C_2^{-1}(\gamma) \\
&\quad + \Delta_\sigma C_2^{-1}(\gamma) A^0 R_2(\gamma) M_1(\gamma) A^{0\top} C_1^{-1}(\gamma) \\
&= \sigma^2 (C_1^{-1}(\gamma) + C_2^{-1}(\gamma)) \\
&\quad + \Delta_\sigma C_1^{-1}(\gamma) A^0 R_1(\gamma) [M_2(\gamma) A^{0\top} C_2^{-1}(\gamma) - M_1(\gamma) A^{0\top} C_1^{-1}(\gamma)] \\
&\quad + \Delta_\sigma C_2^{-1}(\gamma) A^0 R_2(\gamma) [M_1(\gamma) A^{0\top} C_1^{-1}(\gamma) - M_2(\gamma) A^{0\top} C_2^{-1}(\gamma)]. \qquad (2.65)
\end{aligned}$$

In general, $A^0 \in \mathbb{R}^{p\times q}$. Thus, for the just-identified case, i.e. whenever $p = q$, $A^0 \in \mathbb{R}^{p\times p}$. Moreover, since $\Pi^0 \in \mathbb{R}^{q\times p_1}$, $q \geq p_1$, is of full (column) rank by Assumption 2.1.6, A^0 is also of full rank and thus, invertible. Denote by $A^{0^{-1}}$ the inverse of A^0. Hence, it follows that $(A^0 M_i(\gamma) A^{0\top})^{-1} = A^{0\top^{-1}} M_i^{-1}(\gamma) A^{0^{-1}}$. Therefore,

$$M_2(\gamma) A^{0\top} C_2^{-1}(\gamma) - M_1(\gamma) A^{0\top} C_1^{-1}(\gamma) = 0. \qquad (2.66)$$

By equations (2.65)-(2.66), $\tilde{V}(\gamma) = \sigma^2 (C_1^{-1}(\gamma) + C_2^{-1}(\gamma))$. Finally,

$$\tilde{V}^{-1}(\gamma) = \frac{(C_1^{-1}(\gamma) + C_2^{-1}(\gamma))^{-1}}{\sigma^2} = \frac{C_1(\gamma) C^{-1} C_2(\gamma)}{\sigma^2}$$

which yields the asymptotic equivalence of both, sup-LR and sup-Wald tests in the just-identified case under conditional homoskedasticity.

Note that in this setting, $C_1^{-1}(\gamma) A^0 M_1(\gamma) M^{-1} = (A^{0\top})^{-1} M^{-1}$, which implies that:

$$\begin{aligned}
\tilde{\mathscr{E}}(\gamma) &= C_1^{-1}(\gamma) A^0 [\tilde{\mathscr{G}\mathscr{P}}_{\text{mat},1}(\gamma) \Sigma^{1/2} \tilde{\theta}_z^0 - M_1(\gamma) M^{-1} \tilde{\mathscr{G}\mathscr{P}}_{\text{mat}} \Sigma^{1/2} \check{\theta}_z^0] \\
&\quad - C_2^{-1}(\gamma) A^0 [\tilde{\mathscr{G}\mathscr{P}}_{\text{mat},2}(\gamma) \Sigma^{1/2} \tilde{\theta}_z^0 - M_2(\gamma) M^{-1} \tilde{\mathscr{G}\mathscr{P}}_{\text{mat}} \Sigma^{1/2} \check{\theta}_z^0] \\
&= C_1^{-1}(\gamma) A^0 \tilde{\mathscr{G}\mathscr{P}}_{\text{mat},1}(\gamma) \Sigma^{1/2} \tilde{\theta}_z^0 - C_2^{-1}(\gamma) A^0 \tilde{\mathscr{G}\mathscr{P}}_{\text{mat},2}(\gamma) \Sigma^{1/2} \tilde{\theta}_z^0 \\
&= (A^{0\top})^{-1} [M_1^{-1}(\gamma) \tilde{\mathscr{G}\mathscr{P}}_{\text{mat},1}(\gamma) - M_2^{-1}(\gamma) \tilde{\mathscr{G}\mathscr{P}}_{\text{mat},2}(\gamma)] \Sigma^{1/2} \tilde{\theta}_z^0.
\end{aligned}$$

Also,

$$\begin{aligned}
C_2(\gamma) C^{-1} C_1(\gamma) &= A^0 M_2(\gamma) A^{0\top} (A^{0\top})^{-1} M (A^0)^{-1} A^0 M_2(\gamma) A^{0\top} \\
&= A^0 M_2(\gamma) M^{-1} M_1(\gamma) A^{0\top}.
\end{aligned}$$

Therefore, the asymptotic distribution under conditional homoskedasticity and just-identification is:

$$\begin{aligned} J_1(\gamma) = \frac{1}{\sigma^2}(\Sigma^{1/2}\tilde{\theta}_z^0)^\top [M_1^{-1}(\gamma)\tilde{\mathscr{G}\mathscr{P}}_{\text{mat},1}(\gamma) - M_2^{-1}(\gamma)\tilde{\mathscr{G}\mathscr{P}}_{\text{mat},2}(\gamma)]^\top \\ \times [M_2(\gamma)M^{-1}M_1(\gamma)] \\ \times [M_1^{-1}(\gamma)\tilde{\mathscr{G}\mathscr{P}}_{\text{mat},1}(\gamma) - M_2^{-1}(\gamma)\tilde{\mathscr{G}\mathscr{P}}_{\text{mat},2}(\gamma)]\Sigma^{1/2}\tilde{\theta}_z^0. \end{aligned}$$

□

PROOF OF COROLLARY 2.1. First, by Assumption 2.1.4 , Prob$(q_t \le \gamma)$ is continuous in γ. We will replace the sup over the threshold parameter γ by sup over an equivalent value, Prob$(q_t \le \gamma) = \lambda$. To see how this works, note first that $\Gamma \subset \Gamma^0$. Then, Prob$(q_t \le \gamma_{\min}) = 0$ and Prob$(q_t \le \gamma_{\max}) = 1$ in the sample. Suppose now, that Γ can be defined in terms of a cut-off value, say the κ-th quantile, i.e. $\Gamma = [\gamma_\kappa, \gamma_{1-\kappa}]$. Then equivalently, we have Prob$(q_t \le \gamma) = \lambda$ for all $\gamma \in \Gamma$ where λ is uniformly distributed on $\Lambda_\kappa = (\kappa; 1-\kappa)$, i.e $\lambda \sim U(\Lambda_\kappa)$.
Now, by Assumption 2.3, we have that

$$M_1(\gamma_1 \wedge \gamma_2) = \mathbb{E}[x_t x_t^\top \mathbb{1}_{\{q_t \le \gamma_1 \wedge \gamma_2\}}] = \mathbb{E}[x_t x_t^\top]\mathbb{E}[\mathbb{1}_{\{q_t \le \gamma_1 \wedge \gamma_2\}}] = \min\{\lambda_1, \lambda_2\}M. \tag{2.67}$$

This also implies that

$$M_1(\gamma) = \lambda M \tag{2.68a}$$

$$C_1(\gamma) = A^0 M_1(\gamma) A^{0\top} = \lambda A^0 M A^{0\top} = \lambda C \tag{2.68b}$$

$$M_2(\gamma) = (1-\lambda)M \tag{2.68c}$$

$$C_2(\gamma) = A^0 M_2(\gamma) A^{0\top} = (1-\lambda)A^0 M A^{0\top} = (1-\lambda)C. \tag{2.68d}$$

Therefore,

$$\begin{aligned} \tilde{V}_{12}(\gamma) &= -C_1^{-1}(\gamma)\Delta_\sigma A^0 M_1(\gamma) M^{-1} M_2(\gamma) A^{0\top} C_2^{-1}(\gamma) \\ &= -\Delta_\sigma \lambda^{-1} C^{-1} \lambda (1-\lambda) C (1-\lambda)^{-1} C^{-1} \\ &= -\Delta_\sigma C^{-1} \\ \tilde{V}_1(\gamma) &= \lambda^{-1} C^{-1} [\sigma^2 \lambda C - \Delta_\sigma \lambda^2 C] \lambda^{-1} C^{-1} \\ &= \sigma^2 \lambda^{-1} C^{-1} - \Delta_\sigma C^{-1} \\ \tilde{V}_2(\gamma) &= \sigma^2 (1-\lambda)^{-1} C^{-1} - \Delta_\sigma C^{-1} \\ \tilde{V}(\gamma) &= \tilde{V}_1(\gamma) + \tilde{V}_2(\gamma) - \tilde{V}_{12}(\gamma) - \tilde{V}_{12}^\top(\gamma) \\ &= \sigma^2 \lambda^{-1} C^{-1} + \sigma^2 (1-\lambda)^{-1} C^{-1} \\ &= \sigma^2 \frac{C^{-1}}{\lambda(1-\lambda)}, \end{aligned}$$

implying that

$$\sup_{\gamma\in\Gamma} W_T^{2SLS}(\gamma), \sup_{\gamma\in\Gamma} LR_T^{2SLS}(\gamma) \Rightarrow \sup_{\lambda\in\Lambda_\kappa} \frac{\lambda(1-\lambda)}{\sigma^2}\tilde{\mathcal{E}}^\top(\gamma)C\tilde{\mathcal{E}}(\gamma)$$

Hence, in this situation, the sup Wald and sup LR-test are asymptotically equivalent, no matter whether the system is just- or overidentified.
Next, (2.67) implies that – under Assumptions 2.2 and 2.3 – the Gaussian process $\mathcal{GP}_1(\gamma)$ can be restated as

$$\begin{aligned} &\mathcal{GP}_1(\gamma) = (\Sigma^{1/2}\otimes I_q)\tilde{\mathcal{GP}}_1(\gamma) = (\Sigma^{1/2}\otimes M^{1/2})\mathcal{BM}_{q(p_1+1)}(\lambda) \\ \iff\ &\mathcal{GP}_{\text{mat},1}(\gamma) = M^{1/2}\mathcal{BM}_{\text{mat},q(p_1+1)}(\lambda)\Sigma^{1/2} \end{aligned} \tag{2.69}$$

where $\mathcal{BM}_{q(p_1+1)}$ denotes a $q(p_1+1)\times 1$ vector of independent Brownian motions on the unit interval, and $\mathcal{BM}_{\text{mat},q(p_1+1)}(\lambda)$ is the $q\times(p_1+1)$ matrix with $\text{vec}(\mathcal{BM}_{\text{mat},q(p_1+1)}(\lambda)) = \mathcal{BM}_{q(p_1+1)}(\lambda)$. Equation (2.69) in turn implies that $\mathcal{B}_1(\gamma)$ can be rewritten as $\mathcal{B}_1(\lambda)$. Therefore, we obtain

$$\begin{aligned} \tilde{\mathcal{E}}^\top(\gamma)C_2(\gamma)C^{-1}C_1(\gamma)\tilde{\mathcal{E}}(\gamma) &= [C_1^{-1}(\gamma)\mathcal{B}_1(\gamma) - C_2^{-1}(\gamma)\mathcal{B}_2(\gamma)]^\top \\ &\quad\times C_2(\gamma)C^{-1}C_1(\gamma) \\ &\quad\times [C_1^{-1}(\gamma)\mathcal{B}_1(\gamma) - C_2^{-1}(\gamma)\mathcal{B}_2(\gamma)] \\ &= \frac{1}{\lambda(1-\lambda)}[C^{-1}\mathcal{B}_1(\lambda) - \lambda C^{-1}\mathcal{B}_1(1)]^\top \\ &\quad\times C[C^{-1}\mathcal{B}_1(\lambda) - \lambda C^{-1}\mathcal{B}_1(1)] \\ &= \frac{1}{\lambda(1-\lambda)}[C^{-1/2}\mathcal{B}_1(\lambda) - \lambda C^{-1/2}\mathcal{B}_1(1)]^\top \\ &\quad\times [C^{-1/2}\mathcal{B}_1(\lambda) - \lambda C^{-1/2}\mathcal{B}_1(1)]. \end{aligned} \tag{2.70}$$

Next, we show that the term $C^{-1/2}\mathcal{B}_1(\lambda) - \lambda C^{-1/2}\mathcal{B}_1(1) \overset{D}{=} [(\Sigma^{1/2}\tilde{\theta}_z^0)^\top \otimes I_p][\mathcal{BM}_{p(p_1+1)}(\lambda) - \lambda\mathcal{BM}_{p(p_1+1)}(1)]$, where $\mathcal{BM}_{p(p_1+1)}(\lambda)$ collects in a vector the first p out of each q block of elements of $\mathcal{BM}_{q(p_1+1)}(\lambda)$. Because of (2.68a) and (2.69) it follows that

$$\begin{aligned} \mathcal{B}_1(\lambda) &= A^0[\mathcal{GP}_{\text{mat},1}(\gamma)\tilde{\theta}_z^0 - M_1(\gamma)M^{-1}\mathcal{GP}_{\text{mat},1}\check{\theta}_z^0] \\ &= A^0M^{1/2}[\mathcal{BM}_{\text{mat},q(p_1+1)}(\lambda)\Sigma^{1/2}\tilde{\theta}_z^0 - \lambda\mathcal{BM}_{\text{mat},q(p_1+1)}(1)\Sigma^{1/2}\check{\theta}_z^0]. \end{aligned}$$

Furthermore, recall that $C = A^0MA^{0\top}$. Thus:

$$\begin{aligned} C^{-1/2}\mathcal{B}_1(\lambda) &= (A^0MA^{0\top})^{-1/2}A^0M^{1/2}\mathcal{BM}_{\text{mat},q(p_1+1)}(\lambda)\Sigma^{1/2}\tilde{\theta}_z^0 \\ &\quad - \lambda(A^0MA^{0\top})^{-1/2}A^0M^{1/2}\mathcal{BM}_{\text{mat},q(p_1+1)}(1)\Sigma^{1/2}\check{\theta}_z^0. \end{aligned}$$

Note that because $(A^0MA^{0\top})^{-1/2}(A^0MA^{0\top})(A^0MA^{0\top})^{-1/2}$ is a $p\times p$ projection matrix, pre-multiplying with $(A^0MA^{0\top})^{-1/2}A^0M^{1/2}$ is without loss of generality equal in distribution to selecting the first p rows of the q rows of $BM_{\text{mat},q(p_1+1)}(\lambda)$ (this can be seen by writing down the eigenvalue

decomposition of the projection matrix as in Hall et al. (2012), supplemental appendix, page 22-23), yielding

$$\mathscr{BM}_{\text{mat},p(p_1+1)}(\lambda) \overset{\mathscr{D}}{=} (A^0 M A^{0\top})^{-1/2} A^0 M^{1/2} \mathscr{BM}_{\text{mat},q(p_1+1)}(\lambda)$$
$$\mathscr{BM}_{\text{mat},p(p_1+1)}(1) \overset{\mathscr{D}}{=} (A^0 M A^{0\top})^{-1/2} A^0 M^{1/2} \mathscr{BM}_{\text{mat},q(p_1+1)}(1),$$

where $\mathscr{BM}_{p(p_1+1)}(\lambda) = vec(\mathscr{BM}_{\text{mat},p(p_1+1)}(\lambda))$. From the last statement, using the fact that for generic matrices A, B, we have $vec(AB) = (B' \otimes I) vec(A)$,

$$\begin{aligned}
C^{-1/2}\mathscr{B}_1(\lambda) &\overset{\mathscr{D}}{=} \mathscr{BM}_{\text{mat},p(p_1+1)}(\lambda)\Sigma^{1/2}\tilde{\theta}_z^0 - \lambda\mathscr{BM}_{\text{mat},p(p_1+1)}(1)\Sigma^{1/2}\check{\theta}_z^0 \\
\lambda C^{-1/2}\mathscr{B}_1(1) &\overset{\mathscr{D}}{=} \lambda\mathscr{BM}_{\text{mat},p(p_1+1)}(1)\Sigma^{1/2}\tilde{\theta}_z^0 - \lambda\mathscr{BM}_{\text{mat},p(p_1+1)}(1)\Sigma^{1/2}\check{\theta}_z^0 \\
C^{-1/2}\mathscr{B}_1(\lambda) - \lambda C^{-1/2}\mathscr{B}_1(1) &\overset{\mathscr{D}}{=} \mathscr{BM}_{\text{mat},p(p_1+1)}(\lambda)\Sigma^{1/2}\tilde{\theta}_z^0 - \lambda\mathscr{BM}_{\text{mat},p(p_1+1)}(1)\Sigma^{1/2}\tilde{\theta}_z^0 \\
&= [(\Sigma^{1/2}\tilde{\theta}_z^0)^\top \otimes I_p]\,[\mathscr{BM}_{p(p_1+1)}(\lambda) - \lambda\mathscr{BM}_{p(p_1+1)}(1)] \\
&\equiv [(\Sigma^{1/2}\tilde{\theta}_z^0)^\top \otimes I_p]\mathscr{BB}_{p(p_1+1)}(\lambda). \qquad (2.71)
\end{aligned}$$

Using (2.71),

$$\begin{aligned}
\frac{\mathscr{E}^\top(\gamma)C_2(\gamma)C^{-1}C_1(\gamma)\mathscr{E}(\gamma)}{\sigma^2} &\overset{\mathscr{D}}{=} \frac{\left\{[(\Sigma^{1/2}\tilde{\theta}_z^0)^\top \otimes I_p]\mathscr{BB}_{p(p_1+1)}(\lambda)\right\}^\top \left\{[(\Sigma^{1/2}\tilde{\theta}_z^0)^\top \otimes I_p]\mathscr{BB}_{p(p_1+1)}(\lambda)\right\}}{\lambda(1-\lambda)(\Sigma^{1/2}\tilde{\theta}_z^0)^\top(\Sigma^{1/2}\tilde{\theta}_z^0)} \\
&= \frac{\mathscr{BB}^\top_{p(p_1+1)}\left\{[(\Sigma^{1/2}\tilde{\theta}_z^0)[(\Sigma^{1/2}\tilde{\theta}_z^0)^\top(\Sigma^{1/2}\tilde{\theta}_z^0)]^{-1}(\Sigma^{1/2}\tilde{\theta}_z^0)^\top] \otimes I_p\right\}\mathscr{BB}_{p(p_1+1)}}{\lambda(1-\lambda)}.
\end{aligned}$$

Since $F = (\Sigma^{1/2}\tilde{\theta}_z^0)[(\Sigma^{1/2}\tilde{\theta}_z^0)^\top(\Sigma^{1/2}\tilde{\theta}_z^0)]^{-1}(\Sigma^{1/2}\tilde{\theta}_z^0)^\top$ is a projection matrix, as before, pre-multiplying with $F \otimes I_p$ involves, without loss of generality, selecting the first p elements of $\mathscr{BB}_{p(p_1+1)}$, yielding $\mathscr{BB}_p(\lambda)$. Therefore,

$$\frac{\tilde{\mathscr{E}}^\top(\gamma)C_2(\gamma)C^{-1}C_1(\gamma)\tilde{\mathscr{E}}(\gamma)}{\sigma^2} \overset{\mathscr{D}}{=} \frac{\mathscr{BB}_p^\top(\lambda)\mathscr{BB}_p(\lambda)}{\lambda(1-\lambda)},$$

proving the claim. □

Proofs for Section 2.4.4: 2SLS tests and a TFS

Lemma 2.B.5. *Under Assumption 2.1, $T(\hat{\rho}-\rho^0) = \mathscr{O}_p(1)$, $T^{1/2}(\hat{\Pi}_i - \Pi_i^0) = \mathscr{O}_p(1)$, $i = 1,2$ and it holds that the distribution is as if ρ^0 was known:*

$$T^{1/2}\,\text{vec}(\hat{\Pi}_i(\rho^0) - \Pi_i^0) \xrightarrow{\mathscr{D}} \mathscr{N}(0, S_i),$$

where $S_1 = (I_{p_1} \otimes M_1^{-1}(\rho^0))\mathbb{E}[(u_t u_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t \le \rho^0\}}](I_{p_1} \otimes M_1^{-1}(\rho^0))$ and $S_2 = (I_{p_1} \otimes M_2^{-1}(\rho^0))\mathbb{E}[(u_t u_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t > \rho^0\}}](I_{p_1} \otimes M_2^{-1}(\rho^0))$.

PROOF OF LEMMA 2.B.5. The results $T(\hat{\rho}-\rho^0) = \mathscr{O}_p(1)$, $T^{1/2}(\hat{\Pi}_i - \Pi_i^0) = \mathscr{O}_p(1)$, $i = 1,2$ directly follow from Caner and Hansen (2004), Theorems 1 and 2. We will prove the statement for

$T^{1/2}\text{vec}(\hat{\Pi}_1(\rho^0) - \Pi_1^0)$. The proof for $T^{1/2}\text{vec}(\hat{\Pi}_2(\rho^0) - \Pi_2^0)$ is similar and omitted for brevity.

By construction

$$\begin{aligned}\hat{\Pi}_1(\rho^0) &= (X_1^{\rho^0\top} X_1^{\rho^0})^{-1}(X_1^{\rho^0\top} Z) \\ &= (X_1^{\rho^0\top} X_1^{\rho^0})^{-1}(X_1^{\rho^0\top} X_1^{\rho^0}\Pi_1^0 + X_1^{\rho^0\top} X_2^{\rho^0}\Pi_2^0 + X_1^{\rho^0\top} u) \\ &= \Pi_1^0 + (X_1^{\rho^0\top} X_1^{\rho^0})^{-1}(X_1^{\rho^0\top} u_1^{\rho^0})\end{aligned}$$

where the last equality holds because $X_1^{\rho^0\top} X_2^{\rho^0} = \mathbf{0}$. Hence,

$$\begin{aligned}T^{1/2}\text{vec}(\hat{\Pi}_1(\rho^0) - \Pi_1^0) &= \text{vec}\Big((T^{-1}X_1^{\rho^0\top} X_1^{\rho^0})^{-1}(T^{-1/2}X_1^{\rho^0\top} u)\Big) \\ &= (I_{p_1} \otimes (T^{-1}X_1^{\rho^0\top} X_1^{\rho^0})^{-1})\,\text{vec}(T^{1/2}(X_1^{\rho^0\top} u_1^{\rho^0})).\end{aligned}$$

Next, $(T^{-1}X_1^{\rho^0\top} X_1^{\rho^0})^{-1} \xrightarrow{p} M_1^{-1}(\rho^0)$ and by Lemma 2.B.1

$$T^{1/2}\text{vec}(X_1^{\rho\top} u_1^{\rho}) \Rightarrow \mathscr{GP}_1(\rho).$$

Note that $\mathscr{GP}_1(\rho)$ is a zero-mean Gaussian process with covariance function $\mathscr{C}_{\mathscr{GP}}(\rho_1, \rho_2) = \mathbb{E}[(u_t u_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t \le \rho_1 \wedge \rho_2\}}]$. Therefore,

$$T^{1/2}\text{vec}(\hat{\Pi}_1(\rho^0) - \Pi_1^0) \Rightarrow (I_{p_1} \otimes M_1^{-1}(\rho^0))\mathscr{GP}_1(\rho^0).$$

Because $\mathscr{GP}_1(\rho^0)$ denotes the Gaussian process at a particular value ρ^0 it follows that $\mathscr{GP}_1(\rho^0) \sim \mathscr{N}(0, \mathbb{E}[u_t u_t^\top \otimes x_t x_t^\top \mathbb{1}_{\{q_t \le \rho^0\}}])$ and therefore,

$$T^{1/2}\text{vec}(\hat{\Pi}_1(\rho^0) - \Pi_1^0) \xrightarrow{\mathscr{D}} (I_{p_1} \otimes M_1^{-1}(\rho^0))\mathscr{N}(0, \mathbb{E}[u_t u_t^\top \otimes x_t x_t^\top \mathbb{1}_{\{q_t \le \rho^0\}}]),$$

which concludes the proof. □

Lemma 2.B.6. *Suppose Assumption 2.1 holds. Then, under* $\mathbb{H}_0$*,*

$$T^{-1}\hat{W}_1^{\gamma\top}\hat{W}_1^{\gamma} \xrightarrow{p} A_1^0 M_1(\gamma \wedge \rho^0)A_1^{0\top} + A_2^0(M_1(\gamma) - M_1(\gamma \wedge \rho^0))A_2^{0\top} = C_{A,1}(\gamma)$$

and

$$\begin{aligned}T^{-1/2}\hat{W}_1^{\gamma\top}\tilde{\epsilon}_1^{\gamma} \Rightarrow \quad & A_1^0\left[\mathscr{GP}_{\text{mat},1}(\gamma)\tilde{\theta}_z^0 - R_1(\gamma \wedge \rho^0; \rho^0)\mathscr{GP}_{\text{mat},1}(\rho^0)\check{\theta}_z^0\right] \\ & + A_2^0\Big[(\mathscr{GP}_{\text{mat},1}(\gamma) - \mathscr{GP}_{\text{mat},1}(\gamma \wedge \rho^0))\tilde{\theta}_z^0 \\ & \quad - (R_2(\gamma \wedge \rho^0; \rho^0) - R_2(\gamma; \rho^0))\mathscr{GP}_{\text{mat},2}(\rho^0)\check{\theta}_z^0\Big] \\ & = \mathscr{B}_{A,1}(\gamma)\end{aligned}$$

PROOF OF LEMMA 2.B.6. This proof is done in two parts: First, we show the asymptotic behavior of $T^{-1}\hat{W}_1^{\gamma\top}\hat{W}_1^{\gamma}$ and afterwards the asymptotic behavior of $T^{-1/2}\hat{W}_1^{\gamma\top}\tilde{\epsilon}_1^{\gamma}$.
Also, it will be helpful during the proofs to consider three cases: Case (a) assumes that $\gamma < \rho^0$, Case (b) that $\gamma = \rho^0$ and Case (c) that $\gamma > \rho^0$. There are two sub-cases within each case:

- In case (a) it follows that $\gamma < \hat{\rho}$ because $\hat{\rho} = \rho^0 + o_p(1)$ by Lemma 2.B.2 and $\gamma - \rho^0$ is a fixed strictly negative number by construction. This implies two sub-cases: $(a.1)$ with $\gamma < \hat{\rho} \leq \rho^0$ and $(a.2)$ with $\gamma < \rho^0 < \hat{\rho}$.
- In case (b) there are two sub-cases: $(b.1)$ with $\gamma = \rho^0 \leq \hat{\rho}$ and $(b.2)$ with $\hat{\rho} < \gamma = \rho^0$
- In case (c) it follows that $\gamma > \hat{\rho}$ because $\hat{\rho} = \rho^0 + o_p(1)$ by Lemma 2.B.2 and $\gamma - \rho^0$ is a fixed strictly positive number by construction. This implies two sub-cases: $(c.1)$ with $\hat{\rho} \leq \rho^0 < \gamma$ and $(c.2)$ with $\rho^0 < \hat{\rho} < \gamma$.

Claim (i)**.** Starting with case (a), because $\gamma < \hat{\rho}$ for both possible sub-cases, it holds uniformly in γ that

$$\begin{aligned} T^{-1}\hat{W}_1^{\gamma\top}\hat{W}_1^{\gamma} &= \hat{A}_1(T^{-1}X_1^{\gamma\top}X_1^{\gamma})\hat{A}_1^{\top} \\ &= A_1^0(T^{-1}X_1^{\gamma\top}X_1^{\gamma})A_1^{0\top} + o_p(1) \\ &\xrightarrow{p} A_1^0 M_1(\gamma)A_1^{0\top} \end{aligned} \tag{2.72}$$

by Lemma 2.B.2.
In case (b), we first consider sub-case $(b.1)$. Because $\gamma \leq \hat{\rho}$, it holds uniformly in γ that

$$\begin{aligned} T^{-1}\hat{W}_1^{\gamma\top}\hat{W}_1^{\gamma} &= \hat{A}_1(T^{-1}X_1^{\gamma\top}X_1^{\gamma})\hat{A}_1^{\top} \\ &= A_1^0(T^{-1}X_1^{\gamma\top}X_1^{\gamma})A_1^{0\top} + o_p(1) \\ &\xrightarrow{p} A_1^0 M_1(\gamma)A_1^{0\top} \end{aligned} \tag{2.73}$$

by Lemma 2.B.2. In sub-case $(b.2)$ it follows that

$$\begin{aligned} T^{-1}\hat{W}_1^{\gamma\top}\hat{W}_1^{\gamma} &= T^{-1}\hat{W}_1^{\hat{\rho}\top}\hat{W}_1^{\hat{\rho}} + T^{-1}(\hat{W}_1^{\gamma\top}\hat{W}_1^{\gamma} - \hat{W}_1^{\hat{\rho}\top}\hat{W}_1^{\hat{\rho}}) \\ &= \hat{A}_1(T^{-1}X_1^{\hat{\rho}\top}X_1^{\hat{\rho}})\hat{A}_1^{\top} + \hat{A}_2(T^{-1}X_1^{\rho^0\top}X_1^{\rho^0} - T^{-1}X_1^{\hat{\rho}\top}X_1^{\hat{\rho}})\hat{A}_2^{\top}, \end{aligned} \tag{2.74}$$

because $\hat{\rho} < \gamma = \rho^0$. By Lemma 2.B.5 we have that $\hat{\rho} = \rho^0 + \mathcal{O}_p(T^{-1})$ and therefore,

$$\begin{aligned} T^{-1}X_1^{\hat{\rho}\top}X_1^{\hat{\rho}} &= T^{-1}\sum_{t=1}^{T} x_t x_t^{\top}\mathbb{1}_{\{q_t \leq \hat{\rho}\}} \\ &= T^{-1}\sum_{t=1}^{T} x_t x_t^{\top}\mathbb{1}_{\{q_t \leq \rho^0\}} + T^{-1}\sum_{t=1}^{T} x_t x_t^{\top}(\mathbb{1}_{\{q_t \leq \hat{\rho}\}} - \mathbb{1}_{\{q_t \leq \rho^0\}}) \\ &= T^{-1}X_1^{\rho^0\top}X_1^{\rho^0} + \mathcal{O}_p(T^{-1}) \\ &= T^{-1}X_1^{\rho^0\top}X_1^{\rho^0} + o_p(1). \end{aligned} \tag{2.75}$$

So, (2.74), (2.75) and Lemma 2.B.2 imply, uniformly in γ,

$$T^{-1}\hat{W}_1^{\gamma\top}\hat{W}_1^{\gamma} \xrightarrow{p} A_1^0 M_1(\rho^0)A_1^{0\top} = A_1^0 M_1(\gamma)A_1^{0\top}. \tag{2.76}$$

Last, we consider case (c). In sub-case $(c.1)$ we have uniformly in γ that

$$\begin{aligned} T^{-1}\hat{W}_1^{\gamma\top}\hat{W}_1^{\gamma} &= T^{-1}\hat{W}_1^{\hat{\rho}\top}\hat{W}_1^{\hat{\rho}} + T^{-1}(\hat{W}_1^{\rho^0\top}\hat{W}_1^{\rho^0} - \hat{W}_1^{\hat{\rho}\top}\hat{W}_1^{\hat{\rho}}) \\ &\quad + T^{-1}(\hat{W}_1^{\gamma\top}\hat{W}_1^{\gamma} - \hat{W}_1^{\rho^0\top}\hat{W}_1^{\rho^0}) \\ &= \hat{A}_1(T^{-1}X_1^{\hat{\rho}\top}X_1^{\hat{\rho}})\hat{A}_1^{\top} + \hat{A}_2(T^{-1}X_1^{\rho^0\top}X_1^{\rho^0} - T^{-1}X_1^{\hat{\rho}\top}X_1^{\hat{\rho}})\hat{A}_2^{\top} \\ &\quad + \hat{A}_2(T^{-1}X_1^{\gamma\top}X_1^{\gamma} - T^{-1}X_1^{\rho^0\top}X_1^{\rho^0})\hat{A}_2^{\top} \\ &\xrightarrow{p} A_1^0 M_1(\rho^0)A_1^{0\top} + A_2^0(M_1(\gamma) - M_1(\rho^0))A_2^{0\top} \end{aligned} \tag{2.77}$$

by Lemma 2.B.2. In sub-case $(c.2)$ it follows uniformly in γ that

$$\begin{aligned} T^{-1}\hat{W}_1^{\gamma\top}\hat{W}_1^{\gamma} &= T^{-1}\hat{W}_1^{\rho^0\top}\hat{W}_1^{\rho^0} + T^{-1}(\hat{W}_1^{\hat{\rho}\top}\hat{W}_1^{\hat{\rho}} - \hat{W}_1^{\rho^0\top}\hat{W}_1^{\rho^0}) \\ &\quad + T^{-1}(\hat{W}_1^{\gamma\top}\hat{W}_1^{\gamma} - \hat{W}_1^{\hat{\rho}\top}\hat{W}_1^{\hat{\rho}}) \\ &= \hat{A}_1(T^{-1}X_1^{\rho^0\top}X_1^{\rho^0})\hat{A}_1^{\top} + \hat{A}_1(T^{-1}X_1^{\hat{\rho}\top}X_1^{\hat{\rho}} - T^{-1}X_1^{\rho^0\top}X_1^{\rho^0})\hat{A}_1^{\top} \\ &\quad + \hat{A}_2(T^{-1}X_1^{\gamma\top}X_1^{\gamma} - T^{-1}X_1^{\hat{\rho}\top}X_1^{\hat{\rho}})\hat{A}_2^{\top} \\ &\xrightarrow{p} A_1^0 M_1(\rho^0)A_1^{0\top} + A_2^0(M_1(\gamma) - M_1(\rho^0))A_2^{0\top}. \end{aligned} \tag{2.78}$$

Finally, putting results (2.72), (2.73), (2.76)–(2.78) together yields the claim.

Claim (ii)**.** To show this claim, we present a full proof for case (a). Since cases (b) and (c) follow similar reasoning we only state the most important intermediate results to conclude the claim.

Starting with sub-case $(a.1)$ of (a) it holds that

$$\begin{aligned} T^{-1/2}\hat{W}_1^{\gamma\top}\tilde{\epsilon}_1^{\gamma} &= \hat{A}_1(T^{-1/2}X_1^{\gamma\top}\tilde{\epsilon}_1^{\gamma}) \\ &= \hat{A}_1(T^{-1/2}X_1^{\gamma\top}(\epsilon_1^{\gamma} + (Z_1^{\gamma} - \hat{Z}_1^{\gamma})\theta_z^0) \\ &= \hat{A}_1\left[T^{-1/2}X_1^{\gamma\top}(\epsilon_1^{\gamma} + (X_1^{\gamma}\Pi_1^0 + u_1^{\gamma} - X_1^{\gamma}\hat{\Pi}_1)\theta_z^0)\right] \\ &= \hat{A}_1\left[T^{-1/2}X_1^{\gamma\top}s_1^{\gamma} - (T^{-1}X_1^{\gamma\top}X_1^{\gamma})T^{1/2}(\hat{\Pi}_1 - \Pi_1^0)\theta_z^0\right], \end{aligned} \tag{2.79}$$

By Lemma 2.B.1 it follows that $T^{-1/2}X_1^{\gamma\top}s_1^{\gamma} \Rightarrow \mathscr{GP}_{\text{mat},1}(\gamma)\tilde{\theta}_z^0$ uniformly in γ where $\text{vec}(\mathscr{GP}_{\text{mat},1}(\gamma)) = \mathscr{GP}_1(\gamma)$ with $\mathscr{GP}_1(\gamma)$ as in Lemma 2.B.1 and $\tilde{\theta}_z^0 = (1, \theta_z^{0\top})^{\top}$. Moreover, uniformly in γ

$$\begin{aligned} (T^{-1}X_1^{\gamma\top}X_1^{\gamma})T^{1/2}(\hat{\Pi}_1 - \Pi_1^0)\theta_z^0 &= (T^{-1}X_1^{\gamma\top}X_1^{\gamma})(T^{-1}X_1^{\hat{\rho}\top}X_1^{\hat{\rho}})^{-1}(T^{-1/2}X_1^{\hat{\rho}\top}u_1^{\hat{\rho}})\theta_z^0 \\ &\Rightarrow M_1(\gamma)M_1^{-1}(\rho^0)\mathscr{GP}_{\text{mat},1}(\rho^0)\check{\theta}_z^0 \end{aligned}$$

Therefore, (2.79) behaves uniformly in γ as

$$T^{-1/2}\hat{W}_1^{\gamma\top}\tilde{\epsilon}_1^{\gamma} \Rightarrow A_1^0\left[\mathscr{GP}_{\text{mat},1}(\gamma)\tilde{\theta}_z^0 - R_1(\gamma;\rho^0)\mathscr{GP}_{\text{mat},1}(\rho^0)\check{\theta}_z^0\right]. \tag{2.80}$$

As in sub-case $(a.1)$, for sub-case $(a.2)$ it follows that

$$T^{-1/2}\hat{W}_1^{\gamma\top}\tilde{\epsilon}_1^{\gamma} = \hat{A}_1\left[T^{-1/2}X_1^{\gamma\top}s_1^{\gamma} - (T^{-1}X_1^{\gamma\top}X_1^{\gamma})T^{1/2}(\hat{\Pi}_1 - \Pi_1^0)\theta_z^0\right]. \tag{2.81}$$

We now have[23]

$$\hat{\Pi}_1 - \Pi_1^0 = (X_1^{\rho^0\top} X_1^{\rho^0})^{-1}(X_1^{\rho^0\top} u_1^{\rho^0}) + o_p(1)$$

because

$$\begin{aligned}\hat{\Pi}_1 &= (X_1^{\hat{\rho}\top} X_1^{\hat{\rho}})^{-1}(X_1^{\hat{\rho}\top} Z_1^{\hat{\rho}}) \\ &= (X_1^{\hat{\rho}\top} X_1^{\hat{\rho}})^{-1}(X_1^{\rho^0\top} X_1^{\rho^0}\Pi_1^0 + X_1^{\rho^0\top} X_1^{\rho^0}\Pi_2^0 - X_1^{\hat{\rho}\top} X_1^{\hat{\rho}}\Pi_2^0 + X_1^{\hat{\rho}\top} u_1^{\hat{\rho}}) \\ &= \Pi_1^0 + (X_1^{\rho^0\top} X_1^{\rho^0})^{-1}(X_1^{\rho^0\top} u_1^{\rho^0}) + o_p(1) \qquad (2.82)\end{aligned}$$

by Lemma 2.B.5. So, putting (2.81) and (2.82) together yields uniformly in γ that

$$T^{-1/2}\hat{W}_1^{\gamma\top}\tilde{\epsilon}_1^{\gamma} \Rightarrow A_1^0\left[\mathscr{G}\mathscr{P}_{\text{mat},1}(\gamma)\tilde{\theta}_z^0 - R_1(\gamma;\rho^0)\mathscr{G}\mathscr{P}_{\text{mat},1}(\rho^0)\check{\theta}_z^0\right]. \qquad (2.83)$$

For case (b), sub-case $(b.1)$, it follows, as for sub-case $(a.2)$, uniformly in γ that

$$T^{-1/2}\hat{W}_1^{\gamma\top}\tilde{\epsilon}_1^{\gamma} = \hat{A}_1\left[T^{-1/2}X_1^{\gamma\top}s_1^{\gamma} - (T^{-1}X_1^{\gamma\top}X_1^{\gamma})T^{1/2}(\hat{\Pi}_1 - \Pi_1^0)\theta_z^0\right]$$

with

$$\hat{\Pi}_1 - \Pi_1^0 = (X_1^{\rho^0\top} X_1^{\rho^0})^{-1}(X_1^{\rho^0\top} u_1^{\rho^0}) + o_p(1).$$

So, as for sub-case $(a.2)$, uniformly in γ

$$T^{-1/2}\hat{W}_1^{\gamma\top}\tilde{\epsilon}_1^{\gamma} \Rightarrow A_1^0\left[\mathscr{G}\mathscr{P}_{\text{mat},1}(\gamma)\tilde{\theta}_z^0 - R_1(\gamma;\rho^0)\mathscr{G}\mathscr{P}_{\text{mat},1}(\rho^0)\check{\theta}_z^0\right], \qquad (2.84)$$

where $R_1(\gamma;\rho^0) = I_q$ whenever $\gamma = \rho^0$.
For sub-case $(b.2)$ it holds uniformly in γ that

$$\begin{aligned}T^{-1/2}\hat{W}_1^{\gamma\top}\tilde{\epsilon}_1^{\gamma} &= \hat{A}_1\left[T^{-1/2}X_1^{\hat{\rho}\top}s_1^{\hat{\rho}} - (T^{-1}X_1^{\hat{\rho}\top}X_1^{\hat{\rho}})T^{1/2}(\hat{\Pi}_1 - \Pi_1^0)\theta_z^0\right] \\ &\quad + \hat{A}_2\left[T^{-1/2}(X_1^{\rho^0\top}s_1^{\rho^0} - X_1^{\hat{\rho}\top}s_1^{\hat{\rho}}) - T^{-1}(X_1^{\rho^0\top}X_1^{\rho^0} - X_1^{\hat{\rho}\top}X_1^{\hat{\rho}})T^{1/2}(\hat{\Pi}_2 - \Pi_2^0)\theta_z^0\right] \\ &\Rightarrow A_1^0\left[\mathscr{G}\mathscr{P}_{\text{mat},1}(\gamma)\tilde{\theta}_z^0 - \mathscr{G}\mathscr{P}_{\text{mat},1}(\gamma)\check{\theta}_z^0\right] \qquad (2.85)\end{aligned}$$

by Lemmata 2.B.1, 2.B.2 and Equation (2.75).
Last, we show the claim for case (c). In sub-case $(c.1)$ it holds uniformly in γ that

$$\begin{aligned}T^{-1/2}\hat{W}_1^{\gamma\top}\tilde{\epsilon}_1^{\gamma} &= \hat{A}_1\left[T^{-1/2}X_1^{\hat{\rho}\top}s_1^{\hat{\rho}} - (T^{-1}X_1^{\hat{\rho}\top}X_1^{\hat{\rho}})T^{1/2}(\hat{\Pi}_1 - \Pi_1^0)\theta_z^0\right] \\ &\quad + \hat{A}_2\left[T^{-1/2}(X_1^{\rho^0\top}s_1^{\rho^0} - X_1^{\hat{\rho}\top}s_1^{\hat{\rho}}) - T^{-1}(X_1^{\rho^0\top}X_1^{\rho^0} - X_1^{\hat{\rho}\top}X_1^{\hat{\rho}})T^{1/2}(\hat{\Pi}_2 - \Pi_2^0)\theta_z^0\right] \\ &\quad + \hat{A}_2\left[T^{-1/2}(X_1^{\gamma\top}s_1^{\gamma} - X_1^{\rho^0\top}s_1^{\rho^0}) - T^{-1}(X_1^{\gamma\top}X_1^{\gamma} - X_1^{\rho^0\top}X_1^{\rho^0})T^{1/2}(\hat{\Pi}_2 - \Pi_2^0)\theta_z^0\right] \\ &\quad \Rightarrow A_1^0\left[\mathscr{G}\mathscr{P}_{\text{mat},1}(\rho^0)\tilde{\theta}_z^0 - \mathscr{G}\mathscr{P}_{\text{mat},1}(\rho^0)\check{\theta}_z^0\right] \\ &\quad + A_2^0\left[\mathscr{G}\mathscr{P}_{\text{mat},1}(\gamma)\tilde{\theta}_z^0 - \mathscr{G}\mathscr{P}_{\text{mat},1}(\rho^0)\tilde{\theta}_z^0 - (I_q - R_2(\gamma;\rho^0))\mathscr{G}\mathscr{P}_{\text{mat},2}(\rho^0)\check{\theta}_z^0\right], \qquad (2.86)\end{aligned}$$

[23] Note that in sub-case $(a.1)$ we could also write $\hat{\Pi}_1 - \Pi_1^0 = (X_1^{\rho^0\top}X_1^{\rho^0})^{-1}(X_1^{\rho^0\top}u_1^{\rho^0}) + o_p(1)$. However, the composition of the $o_p(1)$-term is different in both cases, as illustrated in (2.82). E.g. in (2.82) also $X_1^{\rho^0\top}X_1^{\rho^0}\Pi_2^0 - X_1^{\hat{\rho}\top}X_1^{\hat{\rho}}\Pi_2^0$ is included in the $o_p(1)$-term, whereas in $(a.1)$ this term completely vanishes already in samples (rather than only asymptotically) because of the relative locations of γ, ρ^0 and $\hat{\rho}$.

where the middle term drops because $T^{-1/2}(X_1^{\rho^0\top}s_1^{\rho^0} - X_1^{\hat{\rho}\top}s_1^{\hat{\rho}}) = o_p(1)$, $T^{-1}(X_1^{\rho^0\top}X_1^{\rho^0} - X_1^{\hat{\rho}\top}X_1^{\hat{\rho}}) = o_p(1)$ and $T^{1/2}(\hat{\Pi}_2 - \Pi_2^0) = \mathscr{O}_p(1)$ by Lemma 2.B.5.
Last, sub-case $(c.2)$ yields uniformly in γ

$$
\begin{aligned}
T^{-1/2}\hat{W}_1^{\gamma\top}\tilde{\epsilon}_1^{\gamma} &= \hat{A}_1\left[T^{-1/2}X_1^{\rho^0\top}s_1^{\rho^0} - (T^{-1}X_1^{\rho^0\top}X_1^{\rho^0})T^{1/2}(\hat{\Pi}_1 - \Pi_1^0)\theta_z^0\right] \\
&\quad + \hat{A}_1\left[T^{-1/2}(X_1^{\hat{\rho}\top}s_1^{\hat{\rho}} - X_1^{\rho^0\top}s_1^{\rho^0}) - T^{-1}(X_1^{\hat{\rho}\top}X_1^{\hat{\rho}} - X_1^{\rho^0\top}X_1^{\rho^0})T^{1/2}(\hat{\Pi}_2 - \Pi_2^0)\theta_z^0\right] \\
&\quad + \hat{A}_2\left[T^{-1/2}(X_1^{\gamma\top}s_1^{\gamma} - X_1^{\hat{\rho}\top}s_1^{\hat{\rho}}) - T^{-1}(X_1^{\gamma\top}X_1^{\gamma} - X_1^{\hat{\rho}\top}X_1^{\hat{\rho}})T^{1/2}(\hat{\Pi}_2 - \Pi_2^0)\theta_z^0\right] \\
&\quad \Rightarrow A_1^0\left[\mathscr{G}\mathscr{P}_{\text{mat},1}(\rho^0)\tilde{\theta}_z^0 - \mathscr{G}\mathscr{P}_{\text{mat},1}(\rho^0)\check{\theta}_z^0\right] \\
&\quad + A_2^0\left[\mathscr{G}\mathscr{P}_{\text{mat},1}(\gamma)\tilde{\theta}_z^0 - \mathscr{G}\mathscr{P}_{\text{mat},1}(\rho^0)\tilde{\theta}_z^0 - (I_q - R_2(\gamma;\rho^0))\mathscr{G}\mathscr{P}_{\text{mat},2}(\rho^0)\check{\theta}_z^0\right], \qquad (2.87)
\end{aligned}
$$

where the middle term drops because $T^{-1/2}(X_1^{\hat{\rho}\top}s_1^{\hat{\rho}} - X_1^{\rho^0\top}s_1^{\rho^0}) = o_p(1)$, $T^{-1}(X_1^{\hat{\rho}\top}X_1^{\hat{\rho}} - X_1^{\rho^0\top}X_1^{\rho^0}) = o_p(1)$ and $T^{1/2}(\hat{\Pi}_2 - \Pi_2^0) = \mathscr{O}_p(1)$ by Lemma 2.B.5.
Finally, putting (2.80), (2.83)–(2.87) together immediately yields the claim. □

Lemma 2.B.7. *Suppose Assumption 2.1 holds and define* $\hat{\theta}^{\gamma} = \text{vec}(\hat{\theta}_1^{\gamma}, \hat{\theta}_2^{\gamma})$*, and* $\bar{\theta}^0 = \text{vec}(\theta^0, \theta^0)$*. Then, under* $\mathbb{H}_0$ *and for a fixed* γ*,*

$$
T^{1/2}(\hat{\theta}^{\gamma} - \bar{\theta}^0) \Rightarrow \mathscr{N}(0, \Sigma_A^{\gamma})
$$

with

$$
\Sigma_A^{\gamma} = \begin{bmatrix} V_{A,1}(\gamma) & V_{A,12}(\gamma) \\ V_{A,12}^{\top}(\gamma) & V_{A,2}(\gamma) \end{bmatrix}
$$

where $V_{A,1}(\gamma), V_{A,2}(\gamma)$ *and* $V_{A,12}(\gamma)$ *are defined in Definition 2.3.*

PROOF OF LEMMA 2.B.7. First, we define the following quantities

$$
\bar{W} = \begin{bmatrix} \hat{W}_1^{\gamma} & \mathbf{0} \\ \mathbf{0} & \hat{W}_2^{\gamma} \end{bmatrix}, \; \bar{Y} = \begin{bmatrix} Y_1^{\gamma} \\ Y_2^{\gamma} \end{bmatrix}, \; \hat{\theta}^{\gamma} = \begin{bmatrix} \hat{\theta}_1^{\gamma} \\ \hat{\theta}_2^{\gamma} \end{bmatrix}.
$$

With this notation, the 2SLS estimator is

$$
\begin{aligned}
\hat{\theta}^{\gamma} &= (\bar{W}^{\top}\bar{W})^{-1}\bar{W}^{\top}\bar{Y} \\
&= \bar{\theta}^0 + (\bar{W}^{\top}\bar{W})^{-1}\bar{W}^{\top}\bar{\tilde{\epsilon}}
\end{aligned}
$$

where:

$$
\bar{\tilde{\epsilon}} = \begin{bmatrix} \tilde{\epsilon}_1^{\gamma} \\ \tilde{\epsilon}_2^{\gamma} \end{bmatrix} = \begin{bmatrix} \epsilon_1^{\gamma} + (Z - \hat{Z})_1^{\gamma}\theta_z^0 \\ \epsilon_2^{\gamma} + (Z - \hat{Z})_2^{\gamma}\theta_z^0 \end{bmatrix}.
$$

Hence, by Lemma 2.B.6 it immediately follows that

$$
T^{1/2}(\hat{\theta}^{\gamma} - \bar{\theta}^0) \Rightarrow \begin{bmatrix} C_{A,1}^{-1}(\gamma)\mathscr{B}_{A,1}(\gamma) \\ C_{A,2}^{-1}(\gamma)\mathscr{B}_{A,2}(\gamma) \end{bmatrix} \sim \mathscr{N}(0, \Sigma_A^{\gamma}).
$$

for fixed γ. Thus, we are left to derive Σ_A^γ. Start with $V_{A,1}(\gamma)$:

$$
\begin{aligned}
\mathrm{Var}[\mathscr{B}_{A,1}(\gamma)] &= \mathrm{Var}[A_1^0 \mathscr{GP}_{\mathrm{mat},1}(\gamma)\tilde{\theta}_z^0 - A_1^0 M_1(\gamma) M_1^{-1}(\rho^0)\mathscr{GP}_{\mathrm{mat},1}(\rho^0)\check{\theta}_z^0] \\
&= \mathrm{Var}[(\tilde{\theta}_z^{0\top} \otimes A_1^0)\mathscr{GP}_1(\gamma)] + \mathrm{Var}[(\check{\theta}_z^{0\top} \otimes [A_1^0 M_1(\gamma) M_1^{-1}(\rho^0)])\mathscr{GP}_1(\rho^0)] \\
&\quad - \mathrm{Cov}[(\tilde{\theta}_z^{0\top} \otimes A_1^0)\mathscr{GP}_1(\gamma), (\check{\theta}_z^{0\top} \otimes [A_1^0 M_1(\gamma) M_1^{-1}(\rho^0)])\mathscr{GP}_1(\rho^0)] \\
&\quad - \mathrm{Cov}[(\check{\theta}_z^{0\top} \otimes [A_1^0 M_1(\gamma) M_1^{-1}(\rho^0)])\mathscr{GP}_1(\rho^0), (\tilde{\theta}_z^{0\top} \otimes A_1^0)\mathscr{GP}_1(\gamma)] \\
&= (\tilde{\theta}_z^{0\top} \otimes A_1^0)\mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t \le \gamma\}}](\tilde{\theta}_z^0 \otimes A_1^{0\top}) \\
&\quad + (\check{\theta}_z^{0\top} \otimes [A_1^0 M_1(\gamma) M_1^{-1}(\rho^0)])\mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t \le \rho^0\}}](\check{\theta}_z^0 \otimes [M_1^{-1}(\rho^0) M_1(\gamma) A_1^{0\top}]) \\
&\quad - (\tilde{\theta}_z^{0\top} \otimes A_1^0)\mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t \le \gamma\}}](\check{\theta}_z^0 \otimes [M_1^{-1}(\rho^0) M_1(\gamma) A_1^{0\top}]) \\
&\quad - (\check{\theta}_z^{0\top} \otimes [A_1^0 M_1(\gamma) M_1^{-1}(\rho^0)])\mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t \le \gamma\}}](\tilde{\theta}_z^0 \otimes A_1^{0\top}) \\
&= A_1^0 \mathbb{E}[x_t x_t^\top (\epsilon_t + u_t^\top \theta_z^0)^2 \mathbb{1}_{\{q_t \le \gamma\}}] A_1^{0\top} \\
&\quad + A_1^0 M_1(\gamma) M_1^{-1}(\rho^0)\mathbb{E}[x_t x_t^\top (u_t^\top \theta_z^0)^2 \mathbb{1}_{\{q_t \le \rho^0\}}] M_1^{-1}(\rho^0) M_1(\gamma) A_1^{0\top} \\
&\quad - A_1^0 \mathbb{E}[x_t x_t^\top (\epsilon_t u_t^\top \theta_z^0 + \theta_z^{0\top} u_t u_t^\top \theta_z^0)\mathbb{1}_{\{q_t \le \gamma\}}] M_1^{-1}(\rho^0) M_1(\gamma) A_1^{0\top} \\
&\quad - A_1^0 M_1(\gamma) M_1^{-1}(\rho^0)\mathbb{E}[x_t x_t^\top (\epsilon_t u_t^\top \theta_z^0 + \theta_z^{0\top} u_t u_t^\top \theta_z^0)\mathbb{1}_{\{q_t \le \gamma\}}] A_1^{0\top} \\
&= A_1^0 \left[H_1(\gamma) + R_1(\gamma;\rho^0) H_1^u(\rho^0) R_1^\top(\gamma;\rho^0) - R_1(\gamma;\rho^0)(H_1^{\epsilon,u}(\gamma) + H_1^u(\gamma)) \right. \\
&\quad \left. - (H_1^{\epsilon,u}(\gamma) + H_1^u(\gamma)) R_1^\top(\gamma;\rho^0) \right] A_1^{0\top} \qquad (2.88)
\end{aligned}
$$

which yields the claim for $\gamma \le \rho^0$ when pre- and post-multiplied with $C_{A,1}^{-1}(\gamma)$.

Next, we consider $\mathrm{Var}[\mathscr{B}_{A,2}(\gamma)]$. First, note that

$$\mathrm{Var}[\mathscr{B}_{A,2}(\gamma)] = \mathrm{Var}[\mathscr{B}_A] + \mathrm{Var}[\mathscr{B}_{A,1}(\gamma)] - \mathrm{Cov}[\mathscr{B}_A, \mathscr{B}_{A,1}(\gamma)] - \mathrm{Cov}[\mathscr{B}_{A,1}(\gamma), \mathscr{B}_A] \qquad (2.89)$$

where $\mathrm{Var}[\mathscr{B}_{A,1}(\gamma)]$ was already derived in Equation (2.88), and $\mathscr{B}_A = \mathscr{B}_A(\gamma_{max}) = A_1^0 \mathscr{GP}_{\mathrm{mat},1}(\rho^0) e_1 + A_2^0 \mathscr{GP}_{\mathrm{mat},2}(\rho^0) e_1$ was defined right before Theorem 2.3 and $e_1 = \tilde{\theta}_z^0 - \check{\theta}_z^0 = (1, 0, \ldots, 0)^\top$. Because

$$\mathrm{Var}[\mathscr{B}_A] = \mathrm{Var}[A_1^0 \mathscr{GP}_{\mathrm{mat},1}(\rho^0) e_1] + \mathrm{Var}[A_2^0 \mathscr{GP}_{\mathrm{mat},2}(\rho^0) e_1] \qquad (2.90)$$

where we used the fact that $\mathrm{Cov}[\mathscr{GP}_1(\rho^0), \mathscr{GP}_2(\rho^0)] = \mathbb{E}[\mathscr{GP}_1(\rho^0)\mathscr{GP}_2^\top(\rho^0)] = \mathbb{E}[\mathscr{GP}_1(\rho^0)\mathscr{GP}_1^\top] - \mathbb{E}[\mathscr{GP}_1(\rho^0)\mathscr{GP}_1^\top(\rho^0)] = \mathbf{0}$. Equation (2.90) thus reads as

$$
\begin{aligned}
\mathrm{Var}[\mathscr{B}_A] &= (e_1^\top \otimes A_1^0)\mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t \le \rho^0\}}](e_1 \otimes A_1^{0\top}) \\
&\quad + (e_1^\top \otimes A_2^0)\mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t > \rho^0\}}](e_1 \otimes A_2^{0\top}) \\
&= A_1^0 \mathbb{E}[x_t x_t^\top \epsilon_t^2 \mathbb{1}_{\{q_t \le \rho^0\}}] A_1^{0\top} + A_2^0 \mathbb{E}[x_t x_t^\top \epsilon_t^2 \mathbb{1}_{\{q_t > \rho^0\}}] A_2^{0\top} \\
&= A_1^0 H_1^\epsilon(\rho^0) A_1^{0\top} + A_2^0 H_2^\epsilon(\rho^0) A_2^{0\top}. \qquad (2.91)
\end{aligned}
$$

From (2.89), we still need to derive $\mathrm{Cov}[\mathscr{B}_A, \mathscr{B}_{A,1}(\gamma)]$:

$$\begin{aligned}
\mathrm{Cov}[\mathscr{B}_A, \mathscr{B}_{A,1}(\gamma)] &= \mathrm{Cov}[A_1^0 \mathscr{G}\mathscr{P}_{\mathrm{mat},1}(\rho^0) e_1 + A_2^0 \mathscr{G}\mathscr{P}_{\mathrm{mat},2}(\rho^0) e_1, \\
&\qquad A_1^0 \mathscr{G}\mathscr{P}_{\mathrm{mat},1}(\gamma)\tilde{\theta}_z^0 - A_1^0 M_1(\gamma) M_1^{-1}(\rho^0) \mathscr{G}\mathscr{P}_{\mathrm{mat},1}(\rho^0)\check{\theta}_z^0] \\
&= \mathrm{Cov}[A_1^0 \mathscr{G}\mathscr{P}_{\mathrm{mat},1}(\rho^0) e_1, A_1^0 \mathscr{G}\mathscr{P}_{\mathrm{mat},1}(\gamma)\tilde{\theta}_z^0] \\
&\quad - \mathrm{Cov}[A_1^0 \mathscr{G}\mathscr{P}_{\mathrm{mat},1}(\rho^0) e_1, A_1^0 M_1(\gamma) M_1^{-1}(\rho^0) \mathscr{G}\mathscr{P}_{\mathrm{mat},1}(\rho^0)\check{\theta}_z^0] \qquad (2.92)
\end{aligned}$$

where the last equality holds since $\gamma \leq \rho^0$ implies that $\mathrm{Cov}[\mathscr{G}\mathscr{P}_1(\gamma), \mathscr{G}\mathscr{P}_2(\rho^0)] = \mathrm{Cov}[\mathscr{G}\mathscr{P}_1(\rho^0), \mathscr{G}\mathscr{P}_2(\rho^0)] = \mathbf{0}$. Thus, Equation (2.92) can then be stated as

$$\begin{aligned}
\mathrm{Cov}[\mathscr{B}_A, \mathscr{B}_{A,1}(\gamma)] &= (e_1^\top \otimes A_1^0)\mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t \leq \gamma\}}](\tilde{\theta}_z^0 \otimes A_1^{0\top}) \\
&\quad - (e_1^\top \otimes A_1^0)\mathbb{E}[(v_t v_t^\top \otimes x_t x_t^\top)\mathbb{1}_{\{q_t \leq \rho^0\}}](\check{\theta}_z^0 \otimes M_1^{-1}(\rho^0) M_1(\gamma) A_1^{0\top}) \\
&= A_1^0 \mathbb{E}[x_t x_t^\top (\epsilon_t^2 + \epsilon_t u_t^\top \theta_z^0)\mathbb{1}_{\{q_t \leq \gamma\}}] A_1^{0\top} \\
&\quad - A_1^0 \mathbb{E}[x_t x_t^\top (\epsilon_t u_t^\top \theta_z^0)\mathbb{1}_{\{q_t \leq \rho^0\}}] M_1^{-1}(\rho^0) M_1(\gamma) A_1^{0\top} \\
&= A_1^0 \left[H_1^\epsilon(\gamma) + H_1^{\epsilon,u}(\gamma) - H_1^{\epsilon,u}(\rho^0) R_1^\top(\gamma;\rho^0) \right] A_1^{0\top}. \qquad (2.93)
\end{aligned}$$

Note that $\mathrm{Cov}[\mathscr{B}_{A,1}(\gamma), \mathscr{B}_A] = \mathrm{Cov}[\mathscr{B}_A, \mathscr{B}_{A,1}(\gamma)]^\top$. Hence, putting (2.88), (2.89), (2.91) and (2.93) together yields

$$\begin{aligned}
\mathrm{Var}[\mathscr{B}_{A,2}(\gamma)] &= A_1^0 H_1^\epsilon(\rho^0) A_1^{0\top} + A_2^0 H_2^\epsilon(\rho^0) A_2^{0\top} \\
&\quad + A_1^0 \left[H_1(\gamma) + R_1(\gamma;\rho^0) H_1^u(\rho^0) R_1^\top(\gamma;\rho^0) - R_1(\gamma;\rho^0)(H_1^{\epsilon,u}(\gamma) + H_1^u(\gamma)) \right. \\
&\quad \left. - (H_1^{\epsilon,u}(\gamma) + H_1^u(\gamma)) R_1^\top(\gamma;\rho^0) \right] A_1^{0\top} \\
&\quad - A_1^0 \left[H_1^\epsilon(\gamma) + H_1^{\epsilon,u}(\gamma) - H_1^{\epsilon,u}(\rho^0) R_1^\top(\gamma;\rho^0) \right] A_1^{0\top} \\
&\quad - A_1^0 \left[H_1^\epsilon(\gamma) + H_1^{\epsilon,u}(\gamma) - R_1(\gamma;\rho^0) H_1^{\epsilon,u}(\rho^0) \right] A_1^{0\top} \\
&= A_2^0 \, H_2^\epsilon(\rho^0) \, A_2^{0\top} \\
&\quad + A_1^0 \left[H_1^\epsilon(\rho^0) + H_1(\gamma) - 2H_1^{\epsilon,u}(\gamma) - 2H_1^\epsilon(\gamma) \right. \\
&\quad + R_1(\gamma;\rho^0) H_1^u(\rho^0) R_1^\top(\gamma;\rho^0) \\
&\quad + R_1(\gamma;\rho^0)[-H_1^{\epsilon,u}(\gamma) - H_1^u(\gamma) + H_1^{\epsilon,u}(\rho^0)] \\
&\quad \left. + [-H_1^{\epsilon,u}(\gamma) - H_1^u(\gamma) + H_1^{\epsilon,u}(\rho^0)] R_1^\top(\gamma;\rho^0) \right] A_1^{0\top} \\
&= A_2^0 \, H_2^\epsilon(\rho^0) \, A_2^{0\top} \\
&\quad + A_1^0 \left[H_1^\epsilon(\rho^0) - H_1^\epsilon(\gamma) + H_1^u(\gamma) \right. \\
&\quad + R_1(\gamma;\rho^0) H_1^u(\rho^0) R_1^\top(\gamma;\rho^0) \\
&\quad + R_1(\gamma;\rho^0)[H_1^{\epsilon,u}(\rho^0) - H_1^{\epsilon,u}(\gamma) - H_1^u(\gamma)] \\
&\quad \left. + [H_1^{\epsilon,u}(\rho^0) - H_1^{\epsilon,u}(\gamma) - H_1^u(\gamma)] R_1^\top(\gamma;\rho^0) \right] A_1^{0\top}.
\end{aligned}$$

Pre- and post-multiplication with $C_{A,2}^{-1}(\gamma)$ then yields the claim when $\gamma \leq \rho^0$. Finally, we derive an expression for:

$$\mathrm{Cov}[\mathscr{B}_{A,1}(\gamma), \mathscr{B}_{A,2}(\gamma)] = \mathrm{Cov}[\mathscr{B}_{A,1}(\gamma), \mathscr{B}_A] - \mathrm{Cov}[\mathscr{B}_{A,1}(\gamma), \mathscr{B}_{A,1}(\gamma)].$$

Using results (2.93) and (2.88) immediately yields:

$$\begin{aligned}
\mathrm{Cov}(\mathscr{B}_{A,1}(\gamma),\mathscr{B}_{A,2}(\gamma)) &= \mathrm{Cov}(\mathscr{B}_{A,1}(\gamma),\mathscr{B}_{A}) - \mathrm{Cov}(\mathscr{B}_{A,1}(\gamma),\mathscr{B}_{A,1}(\gamma)) \\
&= A_1^0\Big[H_1^{\epsilon}(\gamma) + H_1^{\epsilon,u}(\gamma) - R_1(\gamma;\rho^0)H_1^{\epsilon,u}(\gamma)\Big]A_1^{0\top} \\
&- A_1^0\Big[H_1(\gamma) + R_1(\gamma;\rho^0)H_1^{u}(\rho^0)R_1^{\top}(\gamma;\rho^0) - R_1(\gamma;\rho^0)[H_1^{\epsilon,u}(\gamma) + H_1^{u}(\gamma)] \\
&- [H_1^{\epsilon,u}(\gamma) + H_1^{u}(\gamma)]R_1^{\top}(\gamma;\rho^0)]\Big]A_1^{0\top} \\
&= A_1^0\Big[H_1^{\epsilon}(\gamma) + H_1^{\epsilon,u}(\gamma) - H_1(\gamma) - R_1(\gamma;\rho^0)[H_1^{\epsilon,u}(\rho^0) - H_1^{\epsilon,u}(\gamma) - H_1^{u}(\gamma)] \\
&- R_1(\gamma;\rho^0)[H_1^{u}(\rho^0)R_1^{\top}(\gamma;\rho^0) + [H_1^{\epsilon,u}(\gamma) + H_1^{u}(\gamma)]R_1^{\top}(\gamma;\rho^0)]\Big]A_1^{0\top} \\
&= A_1^0\Big[-H_1^{\epsilon,u}(\gamma) - H_1^{u}(\gamma) - R_1(\gamma;\rho^0)[H_1^{\epsilon,u}(\rho^0) - H_1^{\epsilon,u}(\gamma) - H_1^{u}(\gamma)] \\
&- R_1(\gamma;\rho^0)[H_1^{u}(\rho^0)R_1^{\top}(\gamma;\rho^0) + [H_1^{\epsilon,u}(\gamma) + H_1^{u}(\gamma)]R_1^{\top}(\gamma;\rho^0)]\Big]A_1^{0\top} \\
&= -A_1^0\Big[H_1^{\epsilon,u}(\gamma) + H_1^{u}(\gamma) + R_1(\gamma;\rho^0)[H_1^{\epsilon,u}(\rho^0) - H_1^{\epsilon,u}(\gamma) - H_1^{u}(\gamma)] \\
&+ R_1(\gamma;\rho^0)[H_1^{u}(\rho^0)R_1^{\top}(\gamma;\rho^0) - [H_1^{\epsilon,u}(\gamma) + H_1^{u}(\gamma)]R_1^{\top}(\gamma;\rho^0)]\Big]A_1^{0\top}
\end{aligned}$$

Pre-, respectively post-multiplication with $C_{A,1}^{-1}(\gamma)$, respectively $C_{A,2}^{-1}(\gamma)$ yields the claim for $\mathrm{Cov}(\hat{\theta}_1^{\gamma},\hat{\theta}_2^{\gamma})$ when $\gamma \leq \rho^0$.

The case $\gamma > \rho^0$ is derived in a similar fashion and thus omitted for brevity. □

Lemma 2.B.8. *Suppose Assumption 2.1 holds. Then, under $\mathbb{H}_0$ and uniformly in γ and for $i = 1,2$,*

$$(i)\ \hat{H}_i^{\epsilon}(\gamma) \xrightarrow{p} H_i^{\epsilon}(\gamma) \qquad (ii)\ \hat{H}_i^{\epsilon,u}(\gamma) \xrightarrow{p} H_i^{\epsilon,u}(\gamma)$$
$$(iii)\ \hat{H}_i^{u}(\gamma) \xrightarrow{p} H_i^{u}(\gamma) \qquad (iv)\ \hat{H}_i(\gamma) \xrightarrow{p} H_i(\gamma)$$

PROOF OF LEMMA 2.B.8. Claim (i)**:** Let $\tilde{A}$ be the one of the two matrices A_1^0 and A_2^0 with larger Frobenius-norm. Then

$$\begin{aligned}
\|w_t\|_2 &= \|A_1^0 x_t \mathbb{1}_{\{q_t \leq \rho^0\}} + A_2^0 x_t \mathbb{1}_{\{q_t > \rho^0\}} + \bar{u}_t\|_2 \\
&\leq \|A_1^0\|_F \|x_t\|_2 \mathbb{1}_{\{q_t \leq \rho^0\}} + \|A_2^0\|_F \|x_t\|_2 \mathbb{1}_{\{q_t > \rho^0\}} + \|u_t\|_2 \\
&\leq \|\tilde{A}\|_F \|x_t\|_2 + \|u_t\|_2
\end{aligned}$$

Using this expression along the lines of the proof of Lemma 2.B.4 then yields the claim.

To show Claims (ii)–(iv), we consider the three cases, and their sub-cases, of Lemma 2.B.6 again.

Claim (ii)**:** *Case a:* In both sub-cases we obtain

$$\begin{aligned}
T^{-1}\sum_{\mathscr{T}_1(\gamma)} x_t x_t^{\top}(\hat{\epsilon}_t \hat{u}_t^{\top}\hat{\theta}_z) &= T^{-1}\sum_{\mathscr{T}_1(\gamma)} x_t x_t^{\top}(\hat{\epsilon}_t[(\Pi_1^0 - \hat{\Pi}_1)^{\top}x_t + u_t]^{\top}\hat{\theta}_z) \\
&\xrightarrow{p} H_1^{\epsilon,u}(\gamma)
\end{aligned}$$

where convergence follows along the same lines as in the proof of Lemma 2.B.4.

Case b: In sub-case b.1 it holds, as for *Case a*, that

$$T^{-1}\sum_{\mathscr{T}_1(\gamma)} x_t x_t^\top(\hat{\epsilon}_t\hat{u}_t^\top\hat{\theta}_z) = T^{-1}\sum_{\mathscr{T}_1(\gamma)} x_t x_t^\top([\hat{\epsilon}_t(\Pi_1^0-\hat{\Pi}_1)^\top x_t+u_t]^\top\hat{\theta}_z)$$
$$\xrightarrow{p} H_1^{\epsilon,u}(\gamma).$$

In sub-case b.2 it follows that

$$\begin{aligned}T^{-1}\sum_{\mathscr{T}_1(\gamma)} x_t x_t^\top(\hat{\epsilon}_t\hat{u}_t^\top\hat{\theta}_z) &= T^{-1}\sum_{\mathscr{T}_1(\hat{\rho})} x_t x_t^\top(\hat{\epsilon}_t\hat{u}_t^\top\hat{\theta}_z) + T^{-1}\sum_{\mathscr{T}_1(\gamma)\setminus\mathscr{T}_1(\hat{\rho})} x_t x_t^\top(\hat{\epsilon}_t\hat{u}_t^\top\hat{\theta}_z)\\ &= T^{-1}\sum_{\mathscr{T}_1(\hat{\rho})} x_t x_t^\top(\hat{\epsilon}_t[(\Pi_1^0-\hat{\Pi}_1)^\top x_t+u_t]^\top\hat{\theta}_z)\\ &+ T^{-1}\sum_{\mathscr{T}_1(\gamma)\setminus\mathscr{T}_1(\hat{\rho})} x_t x_t^\top(\hat{\epsilon}_t[(\Pi_1^0-\hat{\Pi}_2)^\top x_t+u_t]^\top\hat{\theta}_z)\\ &\xrightarrow{p} H_1^{\epsilon,u}(\gamma),\end{aligned}$$

where the second term converges to 0 in probability since the sum is of order $\mathscr{O}_p(1)$ and $\hat{\rho}\xrightarrow{p}\rho^0$ implying $\mathscr{T}_1(\rho^0)\setminus\mathscr{T}_1(\hat{\rho})\xrightarrow{p}\emptyset$.

Case c: In sub-case c.1 it holds that

$$\begin{aligned}T^{-1}\sum_{\mathscr{T}_1(\gamma)} x_t x_t^\top(\hat{\epsilon}_t\hat{u}_t^\top\hat{\theta}_z) &= T^{-1}\sum_{\mathscr{T}_1(\hat{\rho})} x_t x_t^\top(\hat{\epsilon}_t\hat{u}_t^\top\hat{\theta}_z) + T^{-1}\sum_{\mathscr{T}_1(\rho^0)\setminus\mathscr{T}_1(\hat{\rho})} x_t x_t^\top(\hat{\epsilon}_t\hat{u}_t^\top\hat{\theta}_z)\\ &+ T^{-1}\sum_{\mathscr{T}_1(\gamma)\setminus\mathscr{T}_1(\rho^0)} x_t x_t^\top(\hat{\epsilon}_t\hat{u}_t^\top\hat{\theta}_z)\\ &= T^{-1}\sum_{\mathscr{T}_1(\hat{\rho})} x_t x_t^\top(\hat{\epsilon}_t[(\Pi_1^0-\hat{\Pi}_1)^\top x_t+u_t]^\top\hat{\theta}_z)\\ &+ T^{-1}\sum_{\mathscr{T}_1(\rho^0)\setminus\mathscr{T}_1(\hat{\rho})} x_t x_t^\top(\hat{\epsilon}_t[(\Pi_1^0-\hat{\Pi}_2)^\top x_t+u_t]^\top\hat{\theta}_z)\\ &+ T^{-1}\sum_{\mathscr{T}_1(\gamma)\setminus\mathscr{T}_1(\rho^0)} x_t x_t^\top(\hat{\epsilon}_t[(\Pi_2^0-\hat{\Pi}_2)^\top x_t+u_t]^\top\hat{\theta}_z)\\ &\xrightarrow{p} H_1^{\epsilon,u}(\rho^0)+H_1^{\epsilon,u}(\gamma)-H_1^{\epsilon,u}(\rho^0)=H_1^{\epsilon,u}(\gamma),\end{aligned}$$

where the first and third term converge by similar arguments as in the proof of Lemma 2.B.4. The second term converges to 0 in probability since the sum is of order $\mathscr{O}_p(1)$ and $\hat{\rho}\xrightarrow{p}\rho^0$ implying $\mathscr{T}_1(\rho^0)\setminus\mathscr{T}_1(\hat{\rho})\xrightarrow{p}\emptyset$ (this notation means that the number of elements in the set $\mathscr{T}_1(\rho^0)\setminus\mathscr{T}_1(\hat{\rho})$ is negligible in the limit as $T\to\infty$).

For sub-case $c.2$ it holds that

$$\begin{aligned}
T^{-1}\sum_{\mathscr{T}_1(\gamma)} x_t x_t^\top(\hat{\epsilon}_t\hat{u}_t^\top\hat{\theta}_z) &= T^{-1}\sum_{\mathscr{T}_1(\rho^0)} x_t x_t^\top(\hat{\epsilon}_t\hat{u}_t^\top\hat{\theta}_z) + T^{-1}\sum_{\mathscr{T}_1(\hat{\rho})\setminus\mathscr{T}_1(\rho^0)} x_t x_t^\top(\hat{\epsilon}_t\hat{u}_t^\top\hat{\theta}_z) \\
&\quad + T^{-1}\sum_{\mathscr{T}_1(\gamma)\setminus\mathscr{T}_1(\hat{\rho})} x_t x_t^\top(\hat{\epsilon}_t\hat{u}_t^\top\hat{\theta}_z) \\
&= T^{-1}\sum_{\mathscr{T}_1(\rho^0)} x_t x_t^\top(\hat{\epsilon}_t[(\Pi_1^0-\hat{\Pi}_1)^\top x_t+u_t]^\top\hat{\theta}_z) \\
&\quad + T^{-1}\sum_{\mathscr{T}_1(\hat{\rho})\setminus\mathscr{T}_1(\rho^0)} x_t x_t^\top(\hat{\epsilon}_t[(\Pi_2^0-\hat{\Pi}_1)^\top x_t+u_t]^\top\hat{\theta}_z) \\
&\quad + T^{-1}\sum_{\mathscr{T}_1(\gamma)\setminus\mathscr{T}_1(\hat{\rho})} x_t x_t^\top(\hat{\epsilon}_t[(\Pi_2^0-\hat{\Pi}_2)^\top x_t+u_t]^\top\hat{\theta}_z) \\
&\xrightarrow{p} H_1^{\epsilon,u}(\rho^0)+H_1^{\epsilon,u}(\gamma)-H_1^{\epsilon,u}(\rho^0) = H_1^{\epsilon,u}(\gamma),
\end{aligned}$$

where the first and third term converge by similar arguments as in the proof of Lemma 2.B.4. The second term converges to 0 in probability since the sum is of order $\mathscr{O}_p(1)$ and $\hat{\rho}\xrightarrow{p}\rho^0$ implying $\mathscr{T}_1(\rho^0)\setminus\mathscr{T}_1(\hat{\rho})\xrightarrow{p}\emptyset$.

Claim (iii)**:** *Case a:* In both sub-cases we obtain

$$\begin{aligned}
T^{-1}\sum_{\mathscr{T}_1(\gamma)} x_t x_t^\top(\hat{u}_t^\top\hat{\theta}_z)^2 &= T^{-1}\sum_{\mathscr{T}_1(\gamma)} x_t x_t^\top([(\Pi_1^0-\hat{\Pi}_1)^\top x_t+u_t]^\top\hat{\theta}_z)^2 \\
&\xrightarrow{p} H_1^{u}(\gamma)
\end{aligned}$$

where convergence follows along the same lines as in the proof of Lemma 2.B.4.

Case b: In sub-case $b.1$ it also holds, as before, that

$$\begin{aligned}
T^{-1}\sum_{\mathscr{T}_1(\gamma)} x_t x_t^\top(\hat{u}_t^\top\hat{\theta}_z)^2 &= T^{-1}\sum_{\mathscr{T}_1(\gamma)} x_t x_t^\top([(\Pi_1^0-\hat{\Pi}_1)^\top x_t+u_t]^\top\hat{\theta}_z)^2 \\
&\xrightarrow{p} H_1^{u}(\gamma).
\end{aligned}$$

In sub-case $b.2$ it follows that

$$\begin{aligned}
T^{-1}\sum_{\mathscr{T}_1(\gamma)} x_t x_t^\top(\hat{u}_t^\top\hat{\theta}_z)^2 &= T^{-1}\sum_{\mathscr{T}_1(\hat{\rho})} x_t x_t^\top(\hat{u}_t^\top\hat{\theta}_z)^2 + T^{-1}\sum_{\mathscr{T}_1(\gamma)\setminus\mathscr{T}_1(\hat{\rho})} x_t x_t^\top(\hat{u}_t^\top\hat{\theta}_z)^2 \\
&= T^{-1}\sum_{\mathscr{T}_1(\hat{\rho})} x_t x_t^\top([(\Pi_1^0-\hat{\Pi}_1)^\top x_t+u_t]^\top\hat{\theta}_z)^2 \\
&\quad + T^{-1}\sum_{\mathscr{T}_1(\gamma)\setminus\mathscr{T}_1(\hat{\rho})} x_t x_t^\top([(\Pi_1^0-\hat{\Pi}_2)^\top x_t+u_t]^\top\hat{\theta}_z)^2 \\
&\xrightarrow{p} H_1^{u}(\gamma),
\end{aligned}$$

where the second term converges to 0 in probability since the sum is of order $\mathscr{O}_p(1)$ and $\hat{\rho}\xrightarrow{p}\rho^0$ implying $\mathscr{T}_1(\rho^0)\setminus\mathscr{T}_1(\hat{\rho})\xrightarrow{p}\emptyset$.

Case c: In sub-case c.1 it holds that

$$\begin{aligned}
T^{-1}\sum_{\mathcal{T}_1(\gamma)} x_t x_t^\top(\hat{u}_t^\top\hat{\theta}_z)^2 &= T^{-1}\sum_{\mathcal{T}_1(\hat{\rho})} x_t x_t^\top(\hat{u}_t^\top\hat{\theta}_z)^2 + T^{-1}\sum_{\mathcal{T}_1(\rho^0)\setminus\mathcal{T}_1(\hat{\rho})} x_t x_t^\top(\hat{u}_t^\top\hat{\theta}_z)^2 \\
&\quad + T^{-1}\sum_{\mathcal{T}_1(\gamma)\setminus\mathcal{T}_1(\rho^0)} x_t x_t^\top(\hat{u}_t^\top\hat{\theta}_z)^2 \\
&= T^{-1}\sum_{\mathcal{T}_1(\hat{\rho})} x_t x_t^\top([(\Pi_1^0-\hat{\Pi}_1)^\top x_t+u_t]^\top\hat{\theta}_z)^2 \\
&\quad + T^{-1}\sum_{\mathcal{T}_1(\rho^0)\setminus\mathcal{T}_1(\hat{\rho})} x_t x_t^\top([(\Pi_1^0-\hat{\Pi}_2)^\top x_t+u_t]^\top\hat{\theta}_z)^2 \\
&\quad + T^{-1}\sum_{\mathcal{T}_1(\gamma)\setminus\mathcal{T}_1(\rho^0)} x_t x_t^\top([(\Pi_2^0-\hat{\Pi}_2)^\top x_t+u_t]^\top\hat{\theta}_z)^2 \\
&\xrightarrow{p} H_1^u(\rho^0)+H_1^u(\gamma)-H_1^u(\rho^0)=H_1^u(\gamma),
\end{aligned}$$

where the first and third term converge by similar arguments as in the proof of Lemma 2.B.4. The second term converges to 0 in probability since the sum is of order $\mathcal{O}_p(1)$ and $\hat{\rho}\xrightarrow{p}\rho^0$ implying $\mathcal{T}_1(\rho^0)\setminus\mathcal{T}_1(\hat{\rho})\xrightarrow{p}\emptyset$.
For sub-case c.2 it holds that

$$\begin{aligned}
T^{-1}\sum_{\mathcal{T}_1(\gamma)} x_t x_t^\top(\hat{u}_t^\top\hat{\theta}_z)^2 &= T^{-1}\sum_{\mathcal{T}_1(\rho^0)} x_t x_t^\top(\hat{u}_t^\top\hat{\theta}_z)^2 + T^{-1}\sum_{\mathcal{T}_1(\hat{\rho})\setminus\mathcal{T}_1(\rho^0)} x_t x_t^\top(\hat{u}_t^\top\hat{\theta}_z)^2 \\
&\quad + T^{-1}\sum_{\mathcal{T}_1(\gamma)\setminus\mathcal{T}_1(\hat{\rho})} x_t x_t^\top(\hat{u}_t^\top\hat{\theta}_z)^2 \\
&= T^{-1}\sum_{\mathcal{T}_1(\rho^0)} x_t x_t^\top([(\Pi_1^0-\hat{\Pi}_1)^\top x_t+u_t]^\top\hat{\theta}_z)^2 \\
&\quad + T^{-1}\sum_{\mathcal{T}_1(\hat{\rho})\setminus\mathcal{T}_1(\rho^0)} x_t x_t^\top([(\Pi_2^0-\hat{\Pi}_1)^\top x_t+u_t]^\top\hat{\theta}_z)^2 \\
&\quad + T^{-1}\sum_{\mathcal{T}_1(\gamma)\setminus\mathcal{T}_1(\hat{\rho})} x_t x_t^\top([(\Pi_2^0-\hat{\Pi}_2)^\top x_t+u_t]^\top\hat{\theta}_z)^2 \\
&\xrightarrow{p} H_1^u(\rho^0)+H_1^u(\gamma)-H_1^u(\rho^0)=H_1^u(\gamma),
\end{aligned}$$

where the first and third term converge by similar arguments as in the proof of Lemma 2.B.4. The second term converges to 0 in probability since the sum is of order $\mathcal{O}_p(1)$ and $\hat{\rho}\xrightarrow{p}\rho^0$ implying $\mathcal{T}_1(\rho^0)\setminus\mathcal{T}_1(\hat{\rho})\xrightarrow{p}\emptyset$.

Claim (iv)**:** As in Lemma 2.B.4.

For $i=2$, the proof follows similar steps and omitted for brevity. □

PROOF OF THEOREM 2.3.

(i) sup LR Test: The proof of this result follows the same arguments as in the LFS case. For brevity, we will only display the major differences to the LFS case. As in the LFS case, we split the proof into two parts: in part (i) we will show that $T^{-1}SSR_1(\gamma) \xrightarrow{p} \sigma^2$ and in part (ii) that $SSR_0 - SSR_1(\gamma) \Rightarrow \mathscr{E}_A^\top(\gamma)C_{A,2}(\gamma)C_A^{-1}C_{A,1}(\gamma)\mathscr{E}(\gamma)$.

Part (i)**.** As in the LFS proof (cf. equation (2.48)) it holds uniformly in γ that

$$\begin{aligned}
T^{-1}SSR_1(\gamma) &= T^{-1}[Y_1^\gamma - \hat{W}_1^\gamma\hat{\theta}_1^\gamma]^\top[Y_1^\gamma - \hat{W}_1^\gamma\hat{\theta}_1^\gamma] \\
&\quad + T^{-1}[Y_2^\gamma - \hat{W}_2^\gamma\hat{\theta}_2^\gamma]^\top[Y_2^\gamma - \hat{W}_2^\gamma\hat{\theta}_2^\gamma] \\
&= T^{-1}[\hat{W}_1^\gamma(\theta^0 - \hat{\theta}_1^\gamma) + \tilde{\epsilon}_1^\gamma]^\top[\hat{W}_1^\gamma(\theta^0 - \hat{\theta}_1^\gamma) + \tilde{\epsilon}_1^\gamma] \\
&\quad + T^{-1}[\hat{W}_2^\gamma(\theta^0 - \hat{\theta}_2^\gamma) + \tilde{\epsilon}_2^\gamma]^\top[\hat{W}_2^\gamma(\theta^0 - \hat{\theta}_2^\gamma) + \tilde{\epsilon}_2^\gamma] \\
&= T^{-1}\tilde{\epsilon}^\top\tilde{\epsilon} \\
&\quad + 2(T^{-1}\tilde{\epsilon}_1^\gamma\hat{W}_1^\gamma)(\theta^0 - \hat{\theta}_1^\gamma) + (\theta^0 - \hat{\theta}_1^\gamma)^\top(T^{-1}\hat{W}_1^{\gamma\top}\hat{W}_1^\gamma)(\theta^0 - \hat{\theta}_1^\gamma) \\
&\quad + 2(T^{-1}\tilde{\epsilon}_2^\gamma\hat{W}_2^\gamma)(\theta^0 - \hat{\theta}_2^\gamma) + (\theta^0 - \hat{\theta}_2^\gamma)^\top(T^{-1}\hat{W}_2^{\gamma\top}\hat{W}_2^\gamma)(\theta^0 - \hat{\theta}_2^\gamma) \\
&= T^{-1}\tilde{\epsilon}^\top\tilde{\epsilon} + o_p(1), \qquad (2.94)
\end{aligned}$$

where the last equality holds because, for $i = 1,2$, $T^{-1}\hat{W}_i^{\gamma\top}\tilde{\epsilon}_i^\gamma = o_p(1)$, $T^{-1}\hat{W}_i^{\gamma\top}\hat{W}_i^\gamma = \mathscr{O}_p(1)$ and $\theta^0 - \hat{\theta}_i^\gamma = (T^{-1}\hat{W}_i^{\gamma\top}\hat{W}_i^\gamma)^{-1}(T^{-1}\hat{W}_i^{\gamma\top}\tilde{\epsilon}_i^\gamma) = \mathscr{O}_p(1)o_p(1) = o_p(1)$ uniformly in γ by Lemma 2.B.3. Next, rewrite (2.94) as

$$T^{-1}SSR_1(\gamma) = T^{-1}\tilde{\epsilon}_1^{\rho^0\top}\tilde{\epsilon}_1^{\rho^0} + T^{-1}\tilde{\epsilon}_2^{\rho^0\top}\tilde{\epsilon}_2^{\rho^0} + o_p(1). \qquad (2.95)$$

By construction

$$\tilde{\epsilon}_1^{\rho^0} = \epsilon_1^{\rho^0} + (Z_1^{\rho^0} - \hat{Z}_1^{\rho^0})\theta_z^0$$

and thus

$$\tilde{\epsilon}_1^{\rho^0} = \begin{cases} s_1^{\rho^0} + X_1^{\rho^0}(\Pi_1^0 - \hat{\Pi}_1) & \text{if } \rho^0 \le \hat{\rho} \\ s_1^{\rho^0} + X_1^{\rho^0}(\Pi_1^0 - \hat{\Pi}_1) + o_p(1) & \text{if } \rho^0 > \hat{\rho} \end{cases}$$

where $s_1^{\rho^0} = \epsilon_1^{\rho^0} + u_1^{\rho^0}\theta_z^0$. It can be shown that:

$$\begin{aligned}
T^{-1}\tilde{\epsilon}_1^{\rho^0\top}\tilde{\epsilon}_1^{\rho^0} &= T^{-1}s_1^{\rho^0\top}s_1^{\rho^0} + 2(T^{-1}s_1^{\rho^0\top}X_1^{\rho^0})(\Pi_1^0 - \hat{\Pi}_1) \\
&\quad + (\Pi_1^0 - \hat{\Pi}_1)^\top(T^{-1}X_1^{\rho^0\top}X_1^{\rho^0})(\Pi_1^0 - \hat{\Pi}_1) \\
&= T^{-1}s_1^{\rho^0\top}s_1^{\rho^0} + o_p(1)
\end{aligned}$$

because $T^{-1}s_1^{\rho^0\top}X_1^{\rho^0} = o_p(1)$ and $T^{-1}X_1^{\rho^0\top}X_1^{\rho^0} = \mathscr{O}_p(1)$ and $\Pi_1^0 - \hat{\Pi}_1 = o_p(1)$ by Lemma 2.B.3. Similarly, we obtain

$$T^{-1}\tilde{\epsilon}_2^{\rho^0\top}\tilde{\epsilon}_2^{\rho^0} = T^{-1}s_2^{\rho^0\top}s_2^{\rho^0} + o_p(1).$$

Therefore, (2.95) reads as

$$\begin{aligned}
T^{-1}SSR_1(\gamma) &= T^{-1}s_1^{\rho^0\top}s_1^{\rho^0} + T^{-1}s_2^{\rho^0\top}s_2^{\rho^0} + o_p(1) \\
&= T^{-1}s^\top s + o_p(1) \\
&\xrightarrow{p} \sigma_\epsilon^2 + 2\Sigma_{\epsilon,u}^\top\theta_z^0 + \theta_z^{0\top}\Sigma_u\theta_z^0 \equiv \sigma^2,
\end{aligned}$$

uniformly in γ, proving part (i).

Part (ii)**.** For this part, derivations remain as in the LFS case (up to equation (2.43)). Utilizing Lemma 2.B.4, expressions (2.52) and (2.53) in the LFS proof become

$$T^{-1}\hat{W}_i^{\gamma\top}\hat{W}_i^{\gamma} \xrightarrow{p} C_{A,i}(\gamma)$$

and

$$\hat{\beta} = D_{A,1}(\gamma)\hat{\beta}_1 + D_{A,2}(\gamma)\hat{\beta}_2 + o_p(1)$$

by Lemma 2.B.3 with $D_{A,1}(\gamma) \equiv C_A^{-1}C_{A,1}(\gamma)$ and therefore, $D_{A,2}(\gamma) = C_A^{-1}C_{A,2}(\gamma) = I_p - D_{A,1}(\gamma)$. Consequently, equations (2.52)–(2.54a) in the LFS proof are adjusted in this fashion as well. The following derivations then remain the same.
Last, equation (2.58) from the LFS case now reads as [24]

$$\hat{\beta}_1 - \hat{\beta}_2 = C_{A,1}^{-1}(\gamma)\mathscr{B}_{A,1}(\gamma) - C_{A,2}^{-1}(\gamma)\mathscr{B}_{A,2}(\gamma) \equiv \mathscr{E}_A(\gamma).$$

Thus, as in the LFS case, it follows that

$$\begin{aligned} SSR_0 - SSR_1(\gamma) &= (\hat{\beta}_1 - \hat{\beta}_2)^\top C_{A,2}(\gamma)D_{A,1}(\gamma)(\hat{\beta}_1 - \hat{\beta}_2) + o_p(1) \\ &\Rightarrow \mathscr{E}_A^\top(\gamma)C_{A,2}(\gamma)D_{A,1}(\gamma)\mathscr{E}_A(\gamma) \end{aligned}$$

uniformly in γ. Together with Part (i), $(a.s.)$ continuity of the process $\mathscr{E}_A(\gamma)$, the continuous mapping theorem and weak convergence (uniformly in γ) it then follows that

$$\sup_{\gamma\in\Gamma} \frac{SSR_0 - SSR_1(\gamma)}{SSR_1(\gamma)/T} \Rightarrow \sup_{\gamma\in\Gamma} \frac{\mathscr{E}_A^\top(\gamma)C_{A,2}(\gamma)C_A^{-1}C_{A,1}(\gamma)\mathscr{E}_A(\gamma)}{\sigma^2}$$

proving the claim of the theorem. □

(ii) sup Wald Test:

PROOF. The proof follows the exact same arguments as the proof of Theorem 2.2 by replacing the LFS quantities with the according TFS quantities. □

To write down Corollary 2.B.2 to Theorem 2.3 below, which derives the asymptotic distributions of the 2SLS tests under conditional homoskedasticity, we define the Gaussian processes

$$\tilde{\mathscr{E}}_A(\gamma) = C_{A,1}^{-1}(\gamma)\tilde{\mathscr{B}}_{A,1}(\gamma) - C_{A,2}^{-1}(\gamma)\tilde{\mathscr{B}}_{A,2}(\gamma)$$

and

$$\begin{aligned} \tilde{\mathscr{B}}_{A,1}(\gamma) &= A_1^0\Big[\tilde{\mathscr{G}}\tilde{\mathscr{P}}_{\text{mat},1}(\gamma\wedge\rho^0)\Sigma^{1/2}\tilde{\theta}_z^0 - R_1(\gamma\wedge\rho^0;\rho^0)\tilde{\mathscr{G}}\tilde{\mathscr{P}}_{\text{mat},1}(\rho^0)\Sigma^{1/2}\check{\theta}_z^0\Big] \\ &\quad + A_2^0\Big[\tilde{\mathscr{G}}\tilde{\mathscr{P}}_{\text{mat},1}(\gamma)\Sigma^{1/2}\tilde{\theta}_z^0 - \tilde{\mathscr{G}}\tilde{\mathscr{P}}_{\text{mat},1}(\gamma\wedge\rho^0)\Sigma^{1/2}\tilde{\theta}_z^0\Big] \\ &\quad - A_2^0\Big[(R_2(\gamma\wedge\rho^0;\rho^0) - R_2(\gamma;\rho^0))\tilde{\mathscr{G}}\tilde{\mathscr{P}}_{\text{mat},2}(\rho^0)\Sigma^{1/2}\check{\theta}_z^0\Big] \\ \tilde{\mathscr{B}}_{A,2}(\gamma) &= \tilde{\mathscr{B}}_A(\gamma_{max}) - \tilde{\mathscr{B}}_{A,1}(\gamma), \end{aligned}$$

[24] A^0 is replaced with A_i^0, $i = 1,2$, absorbed in the definition of $\mathscr{B}_{A,1}(\gamma)$.

and $\tilde{\mathscr{GP}}_{\text{mat},1}(\gamma)$ is a $q \times (p_1+1)$ matrix where all columns are independent $q \times 1$ zero mean Gaussian processes with covariance kernel $M_1(\gamma)$.[25] Then we have:

Corollary 2.B.2 (to Theorem 2.3)**.** *Let Z be generated by* (2.2)*, Y be generated by* (2.3)*, and $\hat{Z}$ be calculated by* (2.8)*.Then, under $\mathbb{H}_0$, and Assumptions 2.1, 2.2 and 2.4,*

(i)

$$\sup_{\gamma\in\Gamma} LR^{2SLS}_{T,TFS}(\gamma) \Rightarrow \sup_{\gamma\in\Gamma} \tilde{\mathscr{E}}_A^\top(\gamma) Q_A^{-1}(\gamma) \tilde{\mathscr{E}}_A(\gamma),$$

(ii)

$$\sup_{\gamma\in\Gamma} W^{2SLS}_{T,TFS}(\gamma) \Rightarrow \sup_{\gamma\in\Gamma} \tilde{\mathscr{E}}_A^\top(\gamma) \tilde{V}_A^{-1}(\gamma) \tilde{\mathscr{E}}_A(\gamma).$$

where $\tilde{V}_A(\gamma) = \tilde{V}_{A,1}(\gamma) + \tilde{V}_{A,2}(\gamma) - \tilde{V}_{A,12}(\gamma) - \tilde{V}_{A,12}^\top(\gamma)$ and:

$$\begin{aligned}
\tilde{V}_{A,1}(\gamma) &= C_{A,1}^{-1}(\gamma)\Big[\sigma^2 C_{A,1}(\gamma) - (\sigma^2-\sigma_\epsilon^2) A_1^0 R_1(\gamma;\rho^0) M_1(\gamma) A_1^{0\top}\Big] C_{A,1}^{-1}(\gamma) \\
\tilde{V}_{A,2}(\gamma) &= C_{A,2}^{-1}(\gamma)\Big[\sigma_\epsilon^2 C_{A,2}(\gamma) + (\sigma^2-\sigma_\epsilon^2)(C_{A,1}(\gamma) - A_1^0 R_1(\gamma;\rho^0) M_1(\gamma) A_1^{0\top})\Big] C_{A,2}^{-1}(\gamma) \\
\tilde{V}_{A,12}(\gamma) &= -(\sigma^2-\sigma_\epsilon^2) C_{A,1}^{-1}(\gamma)\Big[C_{A,1}(\gamma) - A_1^0 R_1(\gamma;\rho^0) M_1(\gamma) A_1^{0\top}\Big] C_{A,2}^{-1}(\gamma)
\end{aligned}$$

whenever $\gamma \le \rho^0$. If $\gamma > \rho^0$, then

$$\begin{aligned}
\tilde{V}_{A,1}(\gamma) &= C_{A,1}^{-1}(\gamma)\Big[\sigma_\epsilon^2 C_{A,1}(\gamma) + (\sigma^2-\sigma_\epsilon^2)(C_{A,2}(\gamma) - A_2^0 R_2(\gamma;\rho^0) M_2(\gamma) A_2^{0\top})\Big] C_{A,1}^{-1}(\gamma) \\
\tilde{V}_{A,2}(\gamma) &= C_{A,2}^{-1}(\gamma)\Big[\sigma^2 C_{A,2}(\gamma) - (\sigma^2-\sigma_\epsilon^2) A_2^0 R_2(\gamma;\rho^0) M_2(\gamma) A_2^{0\top}\Big] C_{A,2}^{-1}(\gamma) \\
\tilde{V}_{A,12}(\gamma) &= -(\sigma^2-\sigma_\epsilon^2) C_{A,1}^{-1}(\gamma)\Big[C_{A,2}(\gamma) - A_2^0 R_2(\gamma;\rho^0) M_2(\gamma) A_2^{0\top}\Big] C_{A,2}^{-1}(\gamma)
\end{aligned}$$

Moreover, if the system is just-identified, i.e. if $p = q$, then the two test statistics are asymptotically equivalent with asymptotic distribution given by

$$\sup_{\gamma\in\Gamma} \frac{\tilde{\mathscr{E}}_{\mathscr{A}}^\top(\gamma) C_2(\gamma) C^{-1} C_1(\gamma) \tilde{\mathscr{E}}_{\mathscr{A}}(\gamma)}{\sigma^2}.$$

PROOF OF COROLLARY 2.B.2: The proof follows the exact same arguments as the proof of Corollary 2.B.1. Note that when $p = q$, $\tilde{\mathscr{E}}_{\mathscr{A}}(\gamma)$ does not simplify to $\tilde{\mathscr{E}}(\gamma)$ from the LFS, because ρ^0 does not disappear from the definition of $\tilde{\mathscr{E}}_{\mathscr{A}}(\gamma)$. □

Proofs for Section 2.5: GMM tests

Corollary 2.B.3 (to Theorem 2.4)**.** *Let Z be generated by* (2.1) *and Y be generated by* (2.3)*.Then, under $\mathbb{H}_0$, Assumptions 2.1, 2.2 and $p = q$,*

$$\sup_{\gamma\in\Gamma} W_T^{GMM}(\gamma) \Rightarrow \sup_{\gamma\in\Gamma} J_2(\gamma),$$

[25] Thus, the only difference between the two Gaussian processes $\tilde{\mathscr{GP}}_{\text{mat},1}(\gamma)$ and $\mathscr{GP}_{\text{mat},1}(\gamma)$ lies again in their covariance functions.

where

$$\begin{aligned} J_2(\gamma) = &\left[M_1^{-1}(\gamma)\tilde{\mathscr{G}\mathscr{P}}^{(r1)}_{mat,1}(\gamma) - M_2^{-1}(\gamma)\tilde{\mathscr{G}\mathscr{P}}^{(r1)}_{mat,2}(\gamma)\right]^\top \\ &\times [M_1(\gamma)M^{-1}M_2(\gamma)] \\ &\times \left[M_1^{-1}(\gamma)\tilde{\mathscr{G}\mathscr{P}}^{(r1)}_{mat,1}(\gamma) - M_2^{-1}(\gamma)\tilde{\mathscr{G}\mathscr{P}}^{(r1)}_{mat,2}(\gamma)\right] \end{aligned}$$

and $\tilde{\mathscr{G}\mathscr{P}}^{(r1)}_{mat,i}$ *is the first row of the* $q \times (p_1+1)$ *matrix* $\tilde{\mathscr{G}\mathscr{P}}^{(r2)}_{mat,1}$ *defined in Corollary A2.B.1.*

PROOF OF COROLLARY 2.B.3. We have $H_i^\epsilon(\gamma) = \sigma_\epsilon^2 M_i(\gamma)$, $N_i(\gamma) = A^0 M_i(\gamma)$, $V_{i,GMM} = \sigma_\epsilon^2 (A^0 M_i A^{0\top})^{-1} = (A^{0\top})^{-1} M_i^{-1}(\gamma)(A^0)^{-1}$. The rest follows by plugging these into Theorem 2.4. □

PROOF OF COROLLARY 2.2. As for Corollary 2.1, we can equivalently write $\text{Prob}(q_t \le \gamma) = \lambda$ for all $\gamma \in \Gamma$ where λ is uniformly distributed on $\Lambda_\kappa = (\kappa; 1-\kappa)$, i.e $\lambda \sim U(\Lambda_\kappa)$.
Now, by Assumption 2.3, we have that

$$\begin{aligned} H_1^\epsilon(\gamma) &= \lambda H^\epsilon, \;\; H_2^\epsilon(\gamma) = (1-\lambda)H^\epsilon \\ N_1(\gamma) &= \lambda N, \;\; N_2(\gamma) = (1-\lambda)N \\ V_{GMM,1}(\gamma) &= \lambda^{-1}\left[NH^{\epsilon^{-1}}N^\top\right]^{-1} \\ V_{GMM,2}(\gamma) &= (1-\lambda)^{-1}\left[NH^{\epsilon^{-1}}N^\top\right]^{-1} \end{aligned} \tag{2.96}$$

$$V_{GMM,1}(\gamma) + V_{GMM,2}(\gamma) = \frac{\left[NH^{\epsilon^{-1}}N^\top\right]^{-1}}{\lambda(1-\lambda)} \tag{2.97}$$

$$\begin{aligned} V_{GMM,1}(\gamma)N_1(\gamma)H_1^{\epsilon^{-1}}(\gamma) &= \lambda^{-1}\left[NH^{\epsilon^{-1}}N^\top\right]^{-1}NH^{\epsilon^{-1}} \\ V_{GMM,2}(\gamma)N_2(\gamma)H_2^{\epsilon^{-1}}(\gamma) &= (1-\lambda)^{-1}\left[NH^{\epsilon^{-1}}N^\top\right]^{-1}NH^{\epsilon^{-1}} \end{aligned}$$

Moreover, (2.96) implies that –under Assumptions 2.2 and 2.3– the Gaussian process $\overline{\mathscr{G}\mathscr{P}}_1(\gamma)$ can be restated as

$$\begin{aligned} \overline{\mathscr{G}\mathscr{P}}_1(\gamma) &= H^{\epsilon^{1/2}}\overline{\mathscr{B}\mathscr{M}}(\lambda) \\ \overline{\mathscr{G}\mathscr{P}} &= H^{\epsilon^{1/2}}\overline{\mathscr{B}\mathscr{M}}(1) \end{aligned}$$

where $\overline{\mathscr{B}\mathscr{M}}(\cdot)$ is a $q \times 1$-vector of independent Brownian motions on the unit interval.
Thus, the term $V_{GMM,1}(\gamma)N_1(\gamma)H_1^{\epsilon^{-1}}(\gamma)\overline{\mathscr{G}\mathscr{P}}_1(\gamma) - V_{GMM,2}(\gamma)N_2(\gamma)H_2^{\epsilon^{-1}}(\gamma)\overline{\mathscr{G}\mathscr{P}}_2(\gamma)$ can be restated in terms of λ: as

$$\begin{aligned} &V_{GMM,1}(\gamma)N_1(\gamma)H_1^{\epsilon^{-1}}(\gamma)\overline{\mathscr{G}\mathscr{P}}_1(\gamma) - V_{GMM,2}(\gamma)N_2(\gamma)H_2^{\epsilon^{-1}}(\gamma)\overline{\mathscr{G}\mathscr{P}}_2(\gamma) \\ &\lambda^{-1}\left[NH^{\epsilon^{-1}}N^\top\right]^{-1}NH^{\epsilon^{-1/2}}\overline{\mathscr{B}\mathscr{M}}(\lambda) - (1-\lambda)^{-1}\left[NH^{\epsilon^{-1}}N^\top\right]^{-1}NH^{\epsilon^{-1/2}}(\overline{\mathscr{B}\mathscr{M}}(1) - \overline{\mathscr{B}\mathscr{M}}(\lambda)). \end{aligned} \tag{2.98}$$

Because $\left[NH^{\epsilon^{-1}}N^\top\right]^{-1}NH^{\epsilon^{-1/2}}$ is half of a projection matrix, by similar arguments as for the proof of Corollary 2.1, we obtain the desired result. □

Proofs for Section 3: 2SLS versus GMM estimators

Note: these proofs are provided for simplicity here rather at the beginning of the Appendix because they use results from the proofs above.

Lemma 2.B.9. *Suppose Assumptions 2.1–2.4 hold and that* $p = q = 1$*. Define as in Theorem 2.1,* $\lambda = Prob(q_t \leq \gamma)$*,* $\mu^0 = Prob(q_t \leq \rho^0)$*,* $\alpha = (\mu^0 - \lambda)/(1-\lambda)$*,* $\beta = \mu^0/\lambda$*, and let* $E(x_t^2) = m$*. Then, under* $\mathbb{H}_0$*:*

$$V^*_{1,GMM}(\gamma) = \begin{cases} \frac{\sigma_\epsilon^2}{\lambda m\ \Pi_1^{0^2}} & if\ \gamma \leq \rho^0 \\ \frac{\sigma_\epsilon^2}{\lambda m\ [\beta\Pi_1^0 + (1-\beta)\Pi_2^0]^2} & if\ \gamma > \rho^0 \end{cases}$$

$$V^*_{2,GMM}(\gamma) = \begin{cases} \frac{\sigma_\epsilon^2}{(1-\lambda) m\ [\alpha\Pi_1^0 + (1-\alpha)\Pi_2^0]^2} & if\ \gamma \leq \rho^0 \\ \frac{\sigma_\epsilon^2}{(1-\lambda) m\ \Pi_2^{0^2}} & if\ \gamma > \rho^0. \end{cases}$$

Moreover,

$$V^*_{A,1}(\gamma) = \begin{cases} \frac{\sigma_\epsilon^2}{\lambda m\ \Pi_1^{0^2}} + \frac{\sigma^2 - \sigma_\epsilon^2}{\lambda m\ \Pi_1^{0^2}}\left(1 - \frac{\lambda}{\mu^0}\right) & if\ \gamma \leq \rho^0 \\ \frac{\sigma_\epsilon^2}{\lambda m\ [\beta\Pi_1^{0^2} + (1-\beta)\Pi_2^{0^2}]} + \frac{\Pi_2^{0^2}(1-\lambda)(\sigma^2 - \sigma_\epsilon^2)}{\lambda^2 m\ [\beta\Pi_1^{0^2} + (1-\beta)\Pi_2^{0^2}]^2}\left(1 - \frac{1-\lambda}{1-\mu^0}\right) & if\ \gamma > \rho^0 \end{cases}$$

$$V^*_{A,2}(\gamma) = \begin{cases} \frac{\sigma_\epsilon^2}{(1-\lambda) m\ [\alpha\Pi_1^{0^2} + (1-\alpha)\Pi_2^{0^2}]} + \frac{\Pi_1^{0^2}\lambda(\sigma^2 - \sigma_\epsilon^2)}{(1-\lambda)^2 m\ [\alpha\Pi_1^{0^2} + (1-\alpha)\Pi_2^{0^2}]^2}\left(1 - \frac{\lambda}{\mu^0}\right) & if\ \gamma \leq \rho^0 \\ \frac{\sigma_\epsilon^2}{(1-\lambda) m\ \Pi_2^{0^2}} + \frac{\sigma^2 - \sigma_\epsilon^2}{(1-\lambda) m\ \Pi_2^{0^2}}\left(1 - \frac{1-\lambda}{1-\mu^0}\right) & if\ \gamma > \rho^0 \end{cases}$$

PROOF OF LEMMA 2.B.9. First, we show the claim for the GMM case and afterwards for the 2SLS case.

GMM Variances: Let $\gamma \leq \rho^0$. Then, if Assumptions 2.1–2.4 hold, it follows that $H_1^\epsilon(\gamma) = \mathbb{E}[x_t^2 \epsilon_t^2 \mathbb{1}_{\{q_t \leq \gamma\}}]$ $= \mathbb{E}[\mathbb{1}_{\{q_t \leq \gamma\}}] \cdot \mathbb{E}[x_t^2]\sigma_\epsilon^2 = \lambda\sigma_\epsilon^2 m$, $H_1^\epsilon(\gamma)(\gamma) = \mathbb{E}[x_t^2 \epsilon_t^2 \mathbb{1}_{\{q_t > \gamma\}}] = (1-\lambda)\sigma_\epsilon^2 m$, $N_1(\gamma) = \mathbb{E}[x_t z_t \mathbb{1}_{\{q_t \leq \gamma\}}] = \mathbb{E}[x_t^2 \Pi_1^0 \mathbb{1}_{\{q_t \leq \gamma\}}] =$ $\lambda\Pi_1^0 m$, and $N_2(\gamma) = \mathbb{E}[x_t z_t \mathbb{1}_{\{q_t > \gamma\}}] = \mathbb{E}[x_t^2 \Pi_1^0 (\mathbb{1}_{\{q_t \leq \rho^0\}} - \mathbb{1}_{\{q_t \leq \gamma\}})] + \mathbb{E}[x_t^2 \Pi_2^0 \mathbb{1}_{\{q_t > \rho^0\}}] = (\mu^0 - \lambda)\Pi_1^0 m + (1 -$ $\mu^0)\Pi_2^0 m$. Plugging these results into the expressions for $V_{i,GMM}(\gamma)$ defined just before Theorem 2.4 directly yields the claim. The case $\gamma > \rho^0$ is omitted for brevity but follows similar arguments.

2SLS Variances: Let $\gamma \leq \rho^0$. Then, if Assumptions 2.1–2.4 hold, it follows that $M_1(\gamma) = \mathbb{E}[x_t^2 \mathbb{1}_{\{q_t \leq \gamma\}})] =$ λm, $C_{A,1}(\gamma) = \lambda\Pi_1^{0^2} m$, and also that $\Psi_1(\gamma) \equiv \mathbb{E}[v_t v_t^\top x_t^2 \mathbb{1}_{\{q_t \leq \gamma\}}] = \lambda m \Sigma$. Hence, $(\tilde{\theta}_z^{0\top} \otimes A_1^0)\Psi_1(\gamma)(\tilde{\theta}_z^0 \otimes$ $A_1^{0\top}) = \lambda\Pi_1^{0^2} m \tilde{\theta}_z^{0\top} \Sigma \tilde{\theta}_z^0 = \lambda\Pi_1^{0^2} m\sigma^2$, for example. Similar derivations apply for all the other quanti-

ties in $V_{A,1}(\gamma)$ defined in Definition 2.3. Thus, it follows that

$$\begin{aligned}
V_{A,1}(\gamma) &= \frac{1}{\lambda^2\Pi_1^{0^4}m^2}\Big[\lambda\Pi_1^{0^2}m\tilde{\theta}_z^{0\top}\Sigma\tilde{\theta}_z^0+\frac{\lambda^2}{\mu^0}\Pi_1^{0^2}m\check{\theta}_z^{0\top}\Sigma\check{\theta}_z^0-2\frac{\lambda^2}{\mu^0}\Pi_1^{0^2}m\tilde{\theta}_z^{0\top}\Sigma\check{\theta}_z^0\Big]\\
&= \frac{1}{\Pi_1^{0^2}m}\Big[\frac{\tilde{\theta}_z^{0\top}\Sigma\tilde{\theta}_z^0}{\lambda}+\frac{\check{\theta}_z^{0\top}\Sigma\check{\theta}_z^0}{\mu^0}-2\frac{\tilde{\theta}_z^{0\top}\Sigma\check{\theta}_z^0}{\mu^0}\Big]\\
&= \frac{1}{\Pi_1^{0^2}m}\Big[\frac{\mu^0\sigma_\epsilon^2+2\mu^0\theta_z^0\sigma_{\epsilon,u}+\mu^0\theta_z^{0^2}\sigma_u^2+\lambda\theta_z^{0^2}\sigma_u^2-2\lambda\theta_z^0\sigma_{\epsilon,u}-2\lambda\theta_z^{0^2}\sigma_u^2}{\lambda\mu^0}\Big]\\
&= \frac{1}{\Pi_1^{0^2}m}\left[\frac{\sigma_\epsilon^2}{\lambda}+\theta_z^0(\sigma^2-\sigma_\epsilon^2)\Big(\frac{1}{\lambda}-\frac{1}{\mu^0}\Big)\right]=V_{A,1}^*(\gamma),
\end{aligned}$$

where $\sigma^2-\sigma_\epsilon^2=2\sigma_{\epsilon,u}\theta_z^0+\sigma_u^2(\theta_z^0)^2$, and $\sigma_{\epsilon,u}=\Sigma_{\epsilon,u}$, $\sigma_u^2=\Sigma_u$, proving the claim for $V_{A,1}^*(\gamma)$. Next, we derive the desired result for $V_{A,2}^*(\gamma)$. By the same arguments as above it immediately follows that

$$\begin{aligned}
V_{A,2}(\gamma) &= \frac{1}{[(\mu^0-\lambda)\Pi_1^{0^2}+(1-\mu^0)\Pi_2^0]^2m^2}\\
&\quad\times\Big[\mu^0\Pi_1^{0^2}Me_1^\top\Sigma e_1+(1-\mu^0)\Pi_2^{0^2}Me_1^\top\Sigma e_1+\lambda\Pi_1^{0^2}m\tilde{\theta}_z^{0\top}\Sigma\tilde{\theta}_z^0\\
&\quad+\frac{\lambda^2}{\mu^0}\Pi_1^{0^2}m\check{\theta}_z^{0\top}\Sigma\check{\theta}_z^0-2\frac{\lambda^2}{\mu^0}\Pi_1^{0^2}m\tilde{\theta}_z^{0\top}\Sigma\check{\theta}_z^0-2\frac{\lambda^2}{\mu^0}\Pi_1^{0^2}Me_1^\top\Sigma\tilde{\theta}_z^0\\
&\quad+2\lambda\Pi_1^{0^2}Me_1^\top\Sigma\check{\theta}_z^0\Big]
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{[(\mu^0-\lambda)\Pi_1^{0^2}+(1-\mu^0)\Pi_2^0]^2m}\\
&\times\Big[(1-\mu^0)\Pi_2^{0^2}\sigma_\epsilon^2+\mu^0\Pi_1^{0^2}\sigma_\epsilon^2+\lambda\Pi_1^{0^2}\sigma_\epsilon^2+2\lambda\Pi_1^{0^2}\theta_z^0\sigma_{\epsilon,u}+\lambda\Pi_1^{0^2}\theta_z^{0^2}\sigma_u^2\\
&\quad+\frac{\lambda^2}{\mu^0}\Pi_1^{0^2}\theta_z^{0^2}\sigma_u^2-2\frac{\lambda^2}{\mu^0}\Pi_1^{0^2}\theta_z^0\sigma_{\epsilon,u}-2\frac{\lambda^2}{\mu^0}\Pi_1^{0^2}\theta_z^{0^2}\sigma_u^2-2\lambda\Pi_1^{0^2}\sigma_\epsilon^2-2\lambda\Pi_1^{0^2}\theta_z^0\sigma_{\epsilon,u}\\
&\quad+2\lambda\Pi_1^{0^2}\theta_z^0\sigma_{\epsilon,u}\Big]\\
&= \frac{\sigma_\epsilon^2}{[(\mu^0-\lambda)\Pi_1^{0^2}+(1-\mu^0)\Pi_2^0]m}+\frac{\Pi_1^{0^2}\lambda(1-\frac{\lambda}{\mu^0})\theta_z^0(2\sigma_{\epsilon,u}+\theta_z^0\sigma_u^2)}{[(\mu^0-\lambda)A_1^{0^2}+(1-\mu^0)A_2^0]^2m}=V_{A,2}^*(\gamma)
\end{aligned}$$

proving the claim for $V_{A,2}^*(\gamma)$. By a symmetry argument the claim follows for $\gamma>\rho^0$. □

PROOF OF THEOREM 2.1. Part (i): Limiting distributions. This follows from Caner and Hansen (2004) and Lemma 2.B.9 for GMM and Lemma 2.B.7 and Lemma 2.B.9 for 2SLS.

Part (ii): Variance comparisons for TFS. We only analyze the case $\gamma\le\rho^0$; by symmetry, the claim for $\gamma>\rho^0$ follows. From Lemma 2.B.9 it follows that:

$$V_{1,GMM}^*(\gamma)-V_{A,1}^*(\gamma)=-\frac{1}{\lambda\Pi_1^{0^2}m}\left[(\sigma^2-\sigma_\epsilon^2)\Big(1-\frac{\lambda}{\mu^0}\Big)\right].$$

Hence,

$$V^*_{1,GMM}(\gamma) \geq V^*_{A,1}(\gamma) \iff \sigma^2 \leq \sigma^2_\epsilon.$$

For the second subsample,

$$V^*_{2,GMM}(\gamma) = \frac{\sigma^2_\epsilon}{(1-\lambda)m\ [\alpha\Pi^0_1 + (1-\alpha)\Pi^0_2]^2}$$
$$V^*_{A,2}(\gamma) = \frac{\sigma^2_\epsilon}{(1-\lambda)m\ [\alpha\Pi^{0^2}_1 + (1-\alpha)\Pi^{0^2}_2]} + \frac{\Pi^{0^2}_1 \lambda(1-\frac{\lambda}{\mu^0})(\sigma^2-\sigma^2_\epsilon)}{(1-\lambda)^2 m\ [\alpha\Pi^{0^2}_1 + (1-\alpha)\Pi^{0^2}_2]^2}.$$

From this,

$$\begin{aligned} & V^*_{2,GMM}(\gamma) - V^*_{A,2}(\gamma) \geq 0 \\ \iff & \frac{\sigma^2_\epsilon}{(1-\lambda)m\ [\alpha\Pi^0_1 + (1-\alpha)\Pi^0_2]^2} - \frac{\sigma^2_\epsilon}{(1-\lambda)m\ [\alpha\Pi^{0^2}_1 + (1-\alpha)\Pi^{0^2}_2]} \\ & - \frac{\Pi^{0^2}_1 \lambda(1-\frac{\lambda}{\mu^0})(\sigma^2-\sigma^2_\epsilon)}{(1-\lambda)^2 m\ [\alpha\Pi^{0^2}_1 + (1-\alpha)\Pi^{0^2}_2]^2} \geq 0. \end{aligned}$$

Since $[\alpha\Pi^0_1 + (1-\alpha)\Pi^0_2]^2 - [\alpha\Pi^{0^2}_1 + (1-\alpha)\Pi^{0^2}_2] = -\alpha(1-\alpha)(\Pi^0_1 - \Pi^0_2)^2 \leq 0$,

$$\frac{\sigma^2_\epsilon}{(1-\lambda)m\ [\alpha\Pi^0_1 + (1-\alpha)\Pi^0_2]^2} \geq \frac{\sigma^2_\epsilon}{(1-\lambda)m\ [\alpha\Pi^{0^2}_1 + (1-\alpha)\Pi^{0^2}_2]},$$

implying that a sufficient condition for $V^*_{2,GMM}(\gamma) - V^*_{A,2}(\gamma) \geq 0$ is $\sigma^2 \leq \sigma^2_\epsilon$, the same condition that is necessary and sufficient for $V^*_{1,GMM}(\gamma) - V^*_{A,1}(\gamma) \geq 0$.

Part (iii). Variance comparisons for LFS. Here, $\Pi^0_1 = \Pi^0_2 = \pi^0$, and because there is no threshold in the FS, without loss of generality we let $\rho^0 = \gamma_{max} \iff \mu^0 = 1$, and we calculate the variances from $\gamma \leq \rho^0 = \gamma_{max}$. Plugging these into the results of part (ii), we have:

$$V^*_{1,GMM}(\gamma) - V^*_{A,1}(\gamma) = -\frac{(1-\lambda)(\sigma^2-\sigma^2_\epsilon)}{\lambda\pi^{0^2}m} \geq 0 \iff \sigma^2 \leq \sigma^2_\epsilon$$
$$\begin{aligned} V^*_{2,GMM}(\gamma) - V^*_{A,2}(\gamma) &= \frac{\sigma^2_\epsilon}{(1-\lambda)m\ \pi^{0^2}} - \frac{\sigma^2_\epsilon}{(1-\lambda)m\ \pi^{0^2}} - \frac{\lambda(\sigma^2-\sigma^2_\epsilon)}{(1-\lambda)m\ \pi^{0^2}} \\ &= -\frac{\lambda(\sigma^2-\sigma^2_\epsilon)}{(1-\lambda)m\ \pi^{0^2}} \geq 0 \iff \sigma^2 \leq \sigma^2_\epsilon. \end{aligned}$$

Part (iv). We obtain the claim by plugging in $\gamma = \rho^0$ into the variance expressions of Lemma 2.B.9. □

CHAPTER 3

Estimating Sparse Long-Run Precision Matrices for Linear Multivariate Time Series

This chapter is based on the identically entitled working paper

This chapter proposes a novel estimator for the sparse inverse of the long-run covariance matrix (also known as long-run precision matrix) of a multivariate linear time series. The proposed estimator minimizes the ℓ_1-penalized log-likelihood function of an $i.i.d.$ mean-zero normal random vector. This is possible by reinterpreting the likelihood as a special case of the Bregman-divergence which measures the distance between any positive definite and symmetric matrix and the true long-run covariance matrix of the time series.

I show that the resulting LASSO-type estimator is $T^{b/2}$-consistent with $0 < b < \frac{2}{3}$ under the maintained assumption that the dimension of the multivariate linear process is fixed, a result due to the choice of the sharp origin kernel of Phillips et al. (2007). Moreover, it is shown that the adaptive LASSO-type estimator enjoys the oracle property of Zou (2006). That is, the adaptive LASSO estimator chooses the non-zero entries correctly with probability tending to one and the estimates for these entries have the same distribution as the oracle estimator. An extensive Monte Carlo study indicates that the estimator performs reasonably well in a variety of settings, although it tends to underestimate the degree of sparsity.

3.1 Introduction

Covariance matrices $\mathbf{\Sigma}$ of a random vector $\mathbf{y}_t = (y_{1,t},...,y_{N,t})^\top \in \mathbb{R}^N$, $t = 1,...,T$ and N is fixed as T tends to infinity, and their inverses $\mathbf{C} = \mathbf{\Sigma}^{-1}$, also called precision matrices, constitute one of the cornerstones in statistical analysis and econometric applications can be found in estimation (by generalized method of moments), hypotheses testing (Wald tests), linear and quadratic discriminant analysis, and network modeling, to mention a few. Moreover, outside statistics, these quantities are utilized in a wide range of applications such as portfolio selection (inter alia Ledoit and Wolf, 2003; Talih, 2003), wireless communication (inter alia Li et al., 2003) and genome analysis (inter alia Li and Gui, 2006; Segal et al., 2005). One common feature in these applications is that some entries in $\mathbf{C}$ can be equal to zero, indicating that the i-th and j-th coordinates, $y_{i,t}$ and $y_{j,t}$, of $\mathbf{y}_t$ are conditionally uncorrelated.[1] I shall call such precision matrices sparse as to indicate that they contain entries equal to zero. For example, in case of portfolio choice problems this would imply that the returns of assets i and j do not *directly* affect each other. Such problems, where the zero entries in $\mathbf{C}$ need to be determined are also called covariance selection problems, a terminology dating back to Dempster (1972) who first considered such problems for $i.i.d.$ normal data. However, in many economic and econometric applications, such as portfolio choice, the $i.i.d.$ assumption on the data $\mathbf{y}_t$ is violated and, therefore, needs to be relaxed. Moreover, it is often the case that the objects of interest are not the short-run covariance and precision matrices $\mathbf{\Sigma} = \mathbb{E}[\mathbf{y}_t\mathbf{y}_t^\top]$ and $\mathbf{C} = \mathbf{\Sigma}^{-1}$ of a zero mean random vector $\mathbf{y}_t$ but rather the long-run counterparts $\mathbf{\Sigma}_{LR} = \sum_{h\in\mathbb{Z}} \mathbf{\Gamma}(h)$, $\mathbf{\Gamma}(h) = \mathbb{E}[\mathbf{y}_{t+h}\mathbf{y}_t^\top]$ for all $h \in \mathbb{Z}$, and $\mathbf{C}_{LR} = \mathbf{\Sigma}_{LR}^{-1}$ if existing. The goal of this chapter is estimation of such long-run precision matrices $\mathbf{C}_{LR}$ under the constraint that some entries are equal to zero but the econometrician does not know which. In particular, I propose an estimator for such long-run precision matrices generated from potentially conditionally heteroskedastic linear time series.

Even though the focus of this chapter is on long-run precision matrix estimation for linear time series, it is instructive to first outline some recent developments on estimation of $\mathbf{C}$ in the $i.i.d.$ case. This is due to the fact that in principal, as I will show, the same methodology, albeit with a different motivation and scope, can be employed when considering the estimation of the long-run counterpart $\mathbf{C}_{LR}$.

As shown in Lauritzen (1996), there is a close connection between estimation of sparse $\mathbf{C}$ and estimation of *partial correlation networks* (PCNs). In particular, if one views the coordinates of $\mathbf{y}_t$ as nodes in a network, then a connection between nodes i and j exists if and only if $y_{i,t}$ and $y_{j,t}$ are partially uncorrelated. This statement is equivalent to saying that a connection between these nodes exists if and only if the (i,j)-th entry $c_{i,j}$, $i \neq j$, of $\mathbf{C}$ does not equal zero. Based on this interpretation, Meinshausen and Bühlmann (2006) propose an estimator for $\mathbf{C}$ based on neighborhood selection. That is, for a given coordinate i of $\mathbf{y}_t$ their approach aims to

[1] If $\mathbf{y}_t$ is multivariate normal, then they also are conditionally independent.

consistently identify the subset of all remaining nodes which gives an optimal prediction of the value of node i. If this is done for all nodes, Meinshausen and Bühlmann (2006) then show that $\mathbf{C}$ can consistently be constructed from these N individual regression results. However, this approach is essentially a two step procedure where first the appropriate model is selected and afterwards $\hat{\mathbf{C}}$ is constructed. As shown in Breiman (1996) such methods are usually unstable in the sense that small changes in the available data can alter the estimated outcomes to a large degree and are, therefore, undesirable in practice. To remedy this issue, Yuan and Lin (2007) and Friedman et al. (2008) propose alternative LASSO-type estimators based on normal $i.i.d.$ data. Their approaches aim to simultaneously do the model selection and the estimation based on the ℓ_1-penalized multivariate log-likelihood function for mean zero $i.i.d.$ normal random vectors. By doing so, they obtain direct estimators for a potentially sparse $\mathbf{C}$. Based on these initial proposals, the literature on estimating potentially sparse precision matrices for $i.i.d.$ random vectors extended in different directions, cf. Banerjee et al. (2008), Fan et al. (2009), Lam and Fan (2009), Cai et al. (2011), Ravikumar et al. (2011), Cai et al. (2012), Banerjee and Ghosal (2013), Cai et al. (2014), Cai et al. (2016) and the references therein. For example, these authors relax the normality assumption on the data, refine existing theoretical results or consider different penalty terms. The interested reader is referred to these articles or the recent survey of Fan et al. (2016) for more details about such methods for $i.i.d.$ random vectors.

If one is, as is the case in this chapter, interested or in need of such direct estimators when the data $\mathbf{y}_t$ originate from a time series setting, then the available methods are scarce. In particular, to my best knowledge, Chen et al. (2013) is the only work proposing a direct estimator for potentially sparse $\mathbf{C}$ in a time series setting. However, as in the $i.i.d.$ case, these authors only consider estimation of the short-run precision matrix $\mathbf{C}$ and do not pursue an estimator for the long-run counterpart $\mathbf{C}_{LR}$. Thereby, they neglect the autocorrelation part present in the data. This renders their approach inapplicable in cases where the long-run precision matrix is needed. One, and to my best knowledge the only, possible way at the moment to remedy this issue is the recent proposal of Barigozzi and Brownlees (2017). In particular, these authors utilize a VAR(p)-parametrization of the underlying multivariate process and use the analytic expression of the long-run precision matrix for estimation. However, the same critique as for Meinshausen and Bühlmann (2006) applies in this setting. That is, their approach is a two stage procedure where first an appropriate VAR(p)-model needs to be selected and the long-run precision matrix is computed afterwards. Moreover, due to the construction of their algorithm a computationally costly two-dimensional grid search needs to be performed to obtain the two optimal regularization parameters.[2] Moreover, rather than estimating $N(N+1)/2$ distinct covariance elements, this parametric framework requires estimation of pN^2 (for the p $N \times N$-parameter matrices) plus $N(N+1)/2$ parameters (for the precision matrix of the white noise errors).

[2]One regularization parameter is for the set of VAR-parameters and the second for the precision matrix of the resulting residuals.

In this chapter, I propose a novel non-parametric estimation procedure for the long-run precision matrix $\mathbf{C}_{LR}$ of a linear time series $\mathbf{y}_t$ that overcomes these computational and econometric challenges when the dimension of the process N is fixed. The proposed estimator adapts the graphical LASSO of Friedman et al. (2008) to dependent data by considering a penalized Bregman-divergence based on the negative log-determinant of a symmetric positive definite matrix. This specific choice for the Bregman-divergence yields the same objective function as given by the likelihood function based on multivariate $i.i.d.$ zero-mean normal data, akin to Gaussian QMLE. The only difference to the $i.i.d.$ case lies in the fact that the sample covariance matrix is replaced by an estimator for the long-run covariance. In particular, I make use of the HAC-estimator proposed in Phillips et al. (2007) which makes use of the sharp origin kernel rather than the usual kernels such as the Quadratic Spectral, the Parzen or the Bartlett kernels. Under the standard assumptions for consistent estimation of the long-run covariance matrices of Phillips et al. (2007), I show that the obtained LASSO-type estimator is $T^{b/2}$-consistent, $0 < b < \frac{2}{3}$, and provide its asymptotic distribution. Moreover, I show that the resulting adaptive LASSO-type estimator enjoys the oracle property of Zou (2006). That is, the adaptive LASSO-type estimator is able to distinguish between the true zero and non-zero entries in $\mathbf{C}_{LR}$ with probability tending to one as the sample size increases and the asymptotic distribution of the estimates of the non-zero elements is the same as the asymptotic distribution of the oracle estimator, the estimator for which it is known *a priori* which elements of $\mathbf{C}_{LR}$ are zero and which not. Finally, I show in an extensive Monte Carlo study that the proposed estimator performs reasonably well in samples. Moreover, during the simulations, I found that pre-whitening the data is highly recommended. If one does not pre-whiten the data, the bandwidth parameter determined in the same data dependent fashion as in Andrews (1991) and Newey and West (1994), is highly inaccurate in finite samples leading to highly inaccurate estimators of $\mathbf{C}_{LR}$. Finally, in the Monte Carlo study I also found that the proposed LASSO-type estimators tend to underpenalize in small samples. In particular, all non-zero parameters are detected correctly but parameters with true values equal to zero are estimated to be non-zero. Nevertheless, the proposed estimators are to my knowledge the first estimators that can detect sparsity of the long-run precision matrix. Moreover, I show that they outperform in terms of Frobenius the naïve estimator, which is obtained by inverting the long-run covariance estimator and does not provide any help in detecting the true sparsity structure of the long-run precision matrix.

The remainder of this chapter is organized as follows: Section 3.2 motivates and outlines the proposed estimator and Section 3.3 states asymptotic results. Section 3.4 provides an extensive Monte Carlo study and Section 3.5 concludes. Proofs are deferred to Appendix A and Tables to Appendix B. Note that in the remainder of this chapter I drop the LR-subscript from $\mathbf{C}_{LR}$ and $\mathbf{\Sigma}_{LR}$ for notational convenience. Thus, all covariance and precision matrices showing up from

this point onwards should be understood as long-run versions, unless stated otherwise.

Throughout this chapter I denote matrices by bold capital letters, vectors by bold lower case letters and scalars by lower case letters. For a $N \times N$-matrix $\mathbf{M}$ I denote its Frobenius norm by $\|\mathbf{M}\|_F = (\sum_i \sum_j m_{i,j}^2)^{1/2}$ where $m_{i,j}$ denotes the (i,j)-th entry of $\mathbf{M}$ and for a $N \times 1$-vector $\mathbf{v}$ its Euclidean norm by $\|\mathbf{v}\|_2 = (\sum_i v_i^2)^{1/2}$. Moreover, I denote positive definiteness of a matrix $\mathbf{M}$ by $\mathbf{M} > 0$. The first derivative or gradient of a function $f(*)$ is denoted by $\nabla f(*)$ where $*$ is a generic argument which, depending on the context, can either be a scalar, a vector or a matrix and a set is generally depicted by calligraphic capital letter $\mathscr{S}$ and its complement denoted by $\mathscr{S}^{\complement}$. Finally, $\mathbb{N}$ denotes the natural numbers without 0.

3.2 Methodology

3.2.1 Preliminaries

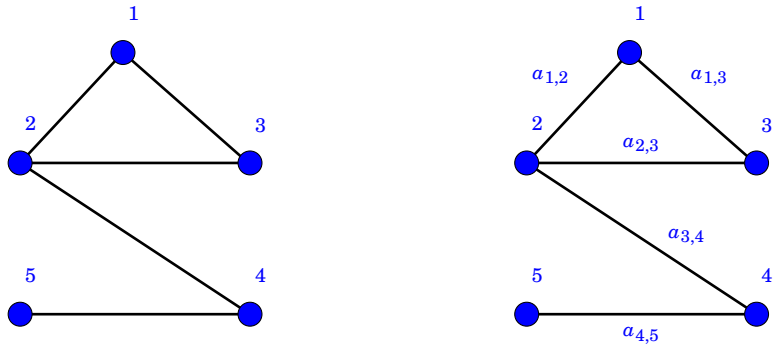
Networks based on Time Series Data As mentioned in the Introduction, the proposed estimator can be used to estimate (weighted) undirected networks based on time series data. However, before I outline how this goal can be achieved I need to briefly introduce the necessary notions on networks and how they can be related to statistical quantities and random data. Based on this knowledge I will then continue to construct an estimator for the problem at hand.

For the purpose of this chapter it suffices to consider undirected networks. An undirected network or graph $\mathscr{G}$ is defined as the tuple $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ where $\mathscr{V} = \{1, 2, ..., N\}$, $N \in \mathbb{N}$, denotes the set of nodes or vertices (individuals, firms or stock returns to give some examples) and $\mathscr{E} \subseteq \mathscr{V} \times \mathscr{V}$ denotes the edge set (connections between nodes). The edge set $\mathscr{E}$ can also be represented by use of an adjacency matrix $\mathbf{A}$. In case of undirected networks we have that $\mathbf{A}$ is symmetric with the diagonal elements being equal to one and the off-diagonal entries having non-zero value $a_{i,j} = a_{j,i}$ if and only if there is an edge between nodes i and j. If it holds that $a_{i,j} = 1$ for all $i \neq j$ then $\mathscr{G}$ is called an unweighted undirected network. In contrast, the network is called weighted undirected if at least two distinct non-zero entries $a_{i,j}$ and $a_{i',j'}$ of the adjacency matrix satisfy $a_{i,j} \neq a_{i',j'}$ and $a_{i,j}$ denotes the weight of the edge between nodes i and j. Note that $a_{i,j} = 1$ is still a possibility. Figure 3.1 illustrates such networks.

Now, one might be interested in the characteristics of the network and these can be derived from the adjacency matrix $\mathbf{A}$. However, in practice it is not always guaranteed that $\mathscr{E}$ or $\mathbf{A}$, which describe the network structure entirely, are known. In such cases, one of them needs to be estimated from random data. For this chapter I follow the vast statistical literature on network analysis and consider partial correlation networks (PCNs). To illustrate the idea, consider the zero-mean $i.i.d.$ random vector $\mathbf{y}_t = (y_{1,t}, ..., y_{N,t})^\top$ with existing covariance matrix $\mathbf{\Sigma}$. Then, in a PCN an edge between nodes i and j exists if and only if the partial correlation coefficient $\rho(i,j)$ between i and j is unequal to zero. Note that nodes i and j in the PCN are represented by the i-th and j-th coordinates of $\mathbf{y}$. Now, it is known that $\rho(i,j)$ is proportional to $c_{i,j}$, cf. Lauritzen

Figure 3.1: Examples of Undirected Networks

The left figure shows an unweighted undirected and the right figure a weighted undirected network as indicated by the weights $a_{i,j}$ next to each edge. Note that in case of an unweighted network no weights are displayed in the figure since they do not provide further information. Nodes are labelled with numerals and lines correspond to edges.



(1996), where $\mathbf{C} = \mathbf{\Sigma}^{-1}$ denotes the precision matrix of $\mathbf{y}$. In particular

$$\rho(i,j) = \frac{c_{i,j}}{\sqrt{c_{i,i}c_{j,j}}} \tag{3.1}$$

Thus, in a PCN an edge between nodes i and j exists if and only if $c_{i,j} \neq 0$ and Equation (3.1) implies that the structure of the PCN can entirely be determined from an estimate of $\mathbf{C}$. Finally, note that for unweighted PCNs we have that $a_{i,j} = a_{j,i} = 1$ if $c_{i,j} = c_{j,i} \neq 0$ and in case of a weighted PCN that $a_{i,j} = a_{j,i} = \rho(i,j)$. As a consequence for PCNs it holds that $\mathscr{E} = \mathscr{S}^{\complement}$ where $\mathscr{S} \equiv \{(i,j) : c_{i,j} = 0\}$ denotes the index set of all zero entries in $\mathbf{C}$.

In case of time series data a PCN can similarly be defined. To make this idea formal, throughout this chapter I assume that $\mathbf{y}_t \in \mathbb{R}^N$, $t = 1..., T$, is a zero-mean $N \times 1$ dimensional linear stochastic process - see Section 3.3 for a precise definition of a linear stochastic process and the Assumptions I impose on this process. I denote by $\mathbf{\Sigma}$ the *long-run covariance matrix* of $\mathbf{y}_t$, i.e. $\mathbf{\Sigma} = \sum_{h \in \mathbb{Z}} \mathbf{\Gamma}(h)$ where $\mathbf{\Gamma}(h) = \mathbb{E}[\mathbf{y}_{t+h}\mathbf{y}_t^\top]$. The PCN of interest in this case is based on $\mathbf{C} = \mathbf{\Sigma}^{-1}$ the *long-run precision matrix* of $\mathbf{y}_t$. For example, if $\mathbf{y}_t$ are asset returns $\mathbf{C}$ can be thought of as the return-network in the long-run equilibrium which describes which assets' stock prices directly influence each other and, for example, how quickly shocks spread through the system. The latter is possible since in a network with a larger number of edges each node can be reached in shorter time since more direct paths from one node to another are present.

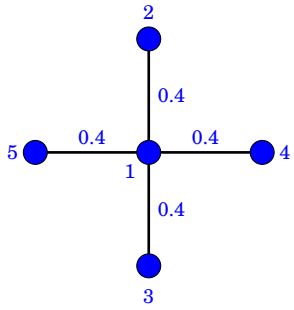
Why not estimating $\mathbf{\Sigma}$ and inverting $\hat{\mathbf{\Sigma}}$? Based on the discussion above one might be tempted to estimate $\mathbf{C}$ by inverting a conventional estimator for $\mathbf{\Sigma}$ in order to obtain the network structure.[3] However, this is not advisable and I will highlight the major reason for that by means

[3] By either using the inverse of the sample covariance for $i.i.d.$ data or the inverse of a HAC estimator for time series data.

of the following example, inspired by El Karoui (2008):

Assume that the underlying PCN is a star-network consisting of 5 nodes, $\mathcal{V}^* = \{1,2,3,4,5\}$. In such a network one of the nodes, w.l.o.g. node 1, is connected to all other nodes and these edges are the only edges within the network, i.e. $\mathcal{E}^* = \{(1,i) : i = 2,3,4,5\} \cup \{(i,1) : i = 2,3,4,5\}$. Moreover, I assume that $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E}^*)$ is weighted with edge weights equal to 0.4 for all edges. This network is illustrated in Figure 3.2.

Figure 3.2: Star-Network $\mathcal{G}^*$



The associated precision and covariance matrices, based on the star-network $\mathcal{G}^*$, are given by[4]

$$\mathbf{C} = \begin{pmatrix} 1 & 0.4 & 0.4 & 0.4 & 0.4 \\ 0.4 & 1 & 0 & 0 & 0 \\ 0.4 & 0 & 1 & 0 & 0 \\ 0.4 & 0 & 0 & 1 & 0 \\ 0.4 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{3.2}$$

respectively

$$\mathbf{\Sigma} = \begin{pmatrix} \frac{25}{9} & -\frac{10}{9} & -\frac{10}{9} & -\frac{10}{9} & -\frac{10}{9} \\ -\frac{10}{9} & \frac{13}{9} & \frac{4}{9} & \frac{4}{9} & \frac{4}{9} \\ -\frac{10}{9} & \frac{4}{9} & \frac{13}{9} & \frac{4}{9} & \frac{4}{9} \\ -\frac{10}{9} & \frac{4}{9} & \frac{4}{9} & \frac{13}{9} & \frac{4}{9} \\ -\frac{10}{9} & \frac{4}{9} & \frac{4}{9} & \frac{4}{9} & \frac{13}{9} \end{pmatrix} \tag{3.3}$$

Now, if $\mathbf{\Sigma}$ would be known *a priori* there is no problem with directly inverting it, as it will yield the correct sparsity structure $\mathcal{S}$. If, however, $\mathbf{\Sigma}$ is unknown it needs to be estimated and in this case the sparsity structure $\mathcal{S}$ of $\mathbf{C}$ is completely or partially removed by the estimation error

[4]Note that the for $i,j \geq 2$ the non-zero off-diagonal elements of $\mathbf{C}$ must satisfy $c_{1,j} = c_{i,1} < \frac{1}{\sqrt{N-1}}$ to ensure positive definiteness of $\mathbf{C}$.

present in $\hat{\mathbf{\Sigma}}$. To see why, I consider the following possible realization of an estimator $\hat{\mathbf{\Sigma}}$ for $\mathbf{\Sigma}$[5]

$$\hat{\mathbf{\Sigma}} = \begin{pmatrix} 3.16 & -1.16 & -1.55 & -0.93 & -0.82 \\ -1.16 & 1.53 & 0.74 & 0.85 & 0.67 \\ -1.55 & 0.74 & 1.67 & 0.43 & 0.32 \\ -0.93 & 0.85 & 0.43 & 1.63 & 0.45 \\ -0.82 & 0.67 & 0.32 & 0.45 & 1.24 \end{pmatrix}. \tag{3.4}$$

Inverting the realized estimator $\hat{\mathbf{\Sigma}}$ yields

$$\hat{\mathbf{C}} = \begin{pmatrix} 0.73 & 0.10 & 0.55 & 0.15 & 0.23 \\ 0.10 & 1.28 & -0.29 & -0.42 & -0.40 \\ 0.55 & -0.29 & 1.17 & 0.10 & 0.18 \\ 0.15 & -0.42 & 0.10 & 0.90 & -0.02 \\ 0.23 & -0.40 & 0.18 & -0.02 & 1.13 \end{pmatrix}. \tag{3.5}$$

As it can be seen, the entire sparsity structure of $\mathbf{C}$ is lost after inverting a realized estimator for $\mathbf{\Sigma}$ in samples since $\hat{\mathscr{S}} = \emptyset \neq \mathscr{S}$. Thus, if one is interested in uncovering the underlying long-run network structure of the linear process $\mathbf{y}_t$ one is well advised to use a direct estimator for $\mathbf{C}$. Proposing such an estimator is precisely the goal of this chapter.

3.2.2 A Bregman-Divergence based Objective Function

To estimate $\mathbf{C}$ under some sparsity constraints I propose a direct estimator for $\mathbf{C}$ based on the Bregman-divergence, cf. Bregman (1967) and also Ravikumar et al. (2011) for a similar application to $i.i.d.$ random vectors. The basic idea behind this approach is to match an estimator $\hat{\mathbf{C}}$ of the long-run precision matrix $\mathbf{C}$ to the long-run covariance matrix $\mathbf{\Sigma}$ so that they are as close as possible. The Bregman-divergence is a suitable measure for that task because it yields convex objective functions which can readily be minimized. For the purpose of this chapter I restrict the analysis to a special case, the log-determinant divergence measure.

Formally, let $b : \mathscr{M} \to \mathbb{R}$ be a continuously differentiable, real valued and strictly convex function defined on the closed convex set $\mathscr{M} = \{\mathbf{M} \in \mathbb{R}^{N,N} : \mathbf{M} = \mathbf{M}^\top, \mathbf{M} \succ 0\}$, the set of all symmetric and positive definite real-valued $N \times N$-matrices. For two points $\mathbf{P}, \mathbf{Q} \in \mathscr{M}$ the Bregman-divergence is then defined as

$$D_b(\mathbf{P}, \mathbf{Q}) = b(\mathbf{P}) - b(\mathbf{Q}) - \langle \nabla b(\mathbf{Q}), \mathbf{P} - \mathbf{Q} \rangle \tag{3.6}$$

[5]For this example, in order to obtain, $\hat{\mathbf{\Sigma}}$ I perturbed $\mathbf{\Sigma}$ by adding a random variable with a uniform distribution on $[-0.5, 0.5]$ to each element on the upper triangle of $\mathbf{\Sigma}$. The new entries are then mirrored onto the lower triangle. For ease of exposition I rounded after the second digits. Inversion of $\hat{\mathbf{\Sigma}}$ is done after rounding the entries of $\hat{\mathbf{\Sigma}}$. The final entries in $\hat{\mathbf{C}}$ are then rounded again after the second digit. This rounding does not take away the main message that by estimating and inverting $\mathbf{\Sigma}$ one loses the sparsity structure of $\mathbf{C}$.

where $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product and $\nabla b(*)$ the gradient of $b(*)$. As mentioned earlier, the function $b(*)$ is chosen such that

$$b(\mathbf{M}) = \begin{cases} -\log\det(\mathbf{M}) & \text{if } \mathbf{M} \in \mathscr{M} \\ \infty & \text{otherwise} \end{cases} \tag{3.7}$$

with $\nabla b(\mathbf{M}) = -\mathbf{M}^{-1}$ (see Section A.4.1 in Boyd and Vandenberghe, 2004) and it is easy to verify that this choice for $b(*)$ satisfies the aforementioned criteria of continuous differentiability, strict convexity and being real valued, cf. Boyd and Vandenberghe (2004).

Next, consider the Bregman-divergence with $b(*)$ as given in (3.7) between the two matrices $\mathbf{\Psi}$ and $\mathbf{C} = \mathbf{\Sigma}^{-1}$. Then, due to symmetry of $\mathbf{C}$, we have that

$$D_b(\mathbf{\Psi}, \mathbf{C}) = -\log\det(\mathbf{\Psi}) + \log\det(\mathbf{C}) + \operatorname{tr}(\mathbf{\Psi}\mathbf{\Sigma} - \mathbf{I}_N) \tag{3.8}$$

since $\langle \nabla b(\mathbf{C}), \mathbf{\Psi} - \mathbf{C} \rangle = \operatorname{tr}(-\mathbf{C}^{-1^\top}(\mathbf{\Psi} - \mathbf{C})) = -\operatorname{tr}(\mathbf{\Sigma}\mathbf{\Psi} - \mathbf{I}_N)$.

Since $D_b(\mathbf{\Psi}, \mathbf{C}) \geq 0$ with equality if and only if $\mathbf{\Psi} = \mathbf{C}$, a natural estimator for $\mathbf{C}$ is given by minimizing

$$-\log\det(\mathbf{\Psi}) + \operatorname{tr}(\mathbf{\Psi}\mathbf{\Sigma}). \tag{3.9}$$

w.r.t. $\mathbf{\Psi} \in \mathscr{M}$. Next, since $\mathbf{\Sigma}$ is unknown in practice it needs to be replaced with a suitable estimator $\hat{\mathbf{\Sigma}}$. Note that I will provide a suitable estimator in Section 3.2.4 and assumptions on it in Section 3.3 below. For now, given a suitable estimator $\hat{\mathbf{\Sigma}}$ for $\mathbf{\Sigma}$ an estimator $\hat{\mathbf{C}}$ for $\mathbf{C}$ is given by

$$\hat{\mathbf{C}} = \hat{\mathbf{\Sigma}}^{-1} = \underset{\mathbf{\Psi} = \mathbf{\Psi}^\top, \mathbf{\Psi} > 0}{\operatorname{argmin}} \; -\log\det(\mathbf{\Psi}) + \operatorname{tr}(\mathbf{\Psi}\hat{\mathbf{\Sigma}}). \tag{3.10}$$

The reader might quickly notice that the objective function in (3.9) coincides with the likelihood function for multivariate $i.i.d.$ Gaussian random vectors and wonder why I formulated it in terms of the Bregman-divergence instead. There are several reasons for this. First, I consider non $i.i.d.$ data which are also allowed to be non-Gaussian. Therefore, it seems odd to me to motivate the choice in a specific $i.i.d.$ Gaussian setting. Second, estimation of long-run precision, respectively covariance matrices is usually not formulated in a parametric likelihood setting but rather in a non-parametric fashion, see the vast literature on HAC estimation. In fact, by using the Bregman-divergence I can reformulate the $i.i.d.$ Gaussian likelihood as a distance measure between two matrices and, therefore, obtain a natural objective function which is free of any distributional and dependency assumptions on the underlying data. This provides a more flexible and intuitive way for deriving an estimator for the long-run precision matrix under a variety of settings.

Moreover, given the generality of the Bregman-divergence an estimator for the long-run precision matrix can be constructed by using different functional forms for $b(*)$ in the hope that the resulting estimator enjoys, for example, faster convergence rates. An example for such a different

form of the Bregman-divergence is the Euclidean distance $\|\operatorname{vec}(\mathbf{P}) - \operatorname{vec}(\mathbf{Q})\|_2^2$, which is implied by setting $b(\mathbf{M}) = \|\operatorname{vec}(\mathbf{M})\|_2^2$. This is not readily possible if the estimator is motivated from the $i.i.d.$ Gaussian likelihood perspective. Therefore, the formulation in terms of the Bregman-divergence also provides interesting directions for future research.

3.2.3 Two LASSO-type Estimators

From (3.10) it is apparent that $\hat{\mathbf{C}} = \hat{\mathbf{\Sigma}}^{-1}$ minimizes the chosen Bregman-divergence. However, as detailed in Section 3.2.1 above this is not a desirable estimator since the sparsity structure $\mathscr{S}$ of $\mathbf{C}$ is lost in finite samples. To rectify this issue I penalize (3.9) by an ℓ_1-penalty on the off-diagonal elements $\psi_{i,j}$, $i \neq j$, of $\mathbf{\Psi}$ since it is known that such a penalty enforces sparsity, if present.[6] That is, after imposing an ℓ_1-penalty on $\psi_{i,j}$, $i \neq j$, I obtain the following objective function

$$q_\lambda^L(\mathbf{\Psi}) = -\log\det(\mathbf{\Psi}) + \operatorname{tr}(\mathbf{\Psi}\hat{\mathbf{\Sigma}}) + \lambda_T \sum_{i \neq j} |\psi_{i,j}| \tag{3.11}$$

A closer inspection reveals that the obtained objective function in (3.11) coincides with the graphical LASSO of Friedman et al. (2008). However, they developed the graphical LASSO estimate the sparse precision matrix for $i.i.d.$ zero mean normal random vectors. Thus, by use of an appropriate function $b(*)$ for the Bregman-divergence, estimating the long-run precision matrix under sparsity constraints poses a similar problem as estimating that of $i.i.d.$ multivariate normal random vectors. This has the advantage that existing efficient algorithms to solve

$$\hat{\mathbf{C}}^L = \underset{\mathbf{\Psi}=\mathbf{\Psi}^\top, \mathbf{\Psi}>0}{\operatorname{argmin}} \; q_\lambda^L(\mathbf{\Psi}) \tag{3.12}$$

can be utilized. These considerations further motivated the particular function choice for the Bregman-divergence.

It is well known that LASSO-type estimators do not always consistently identify the true set of the paramters with value equal to 0, cf. Zou (2006). To circumvent this issue, I follow Zou (2006) and consider an adaptive LASSO-type estimator where the penalty term in (3.11) is augmented by a data dependent weight for each off-diagonal element of the precision matrix. In particular, I consider the adaptive LASSO-type objective function

$$q_\lambda^{aL}(\mathbf{\Psi}) = -\log\det(\mathbf{\Psi}) + \operatorname{tr}(\mathbf{\Psi}\hat{\mathbf{\Sigma}}) + \lambda_T \sum_{i \neq j} \frac{|\psi_{i,j}|}{|\tilde{c}_{i,j}|} \tag{3.13}$$

where $\tilde{\mathbf{C}} = (\tilde{c}_{i,j})$ denotes a pre-estimator for $\mathbf{C}$. Note that I will provide conditions which $\tilde{\mathbf{C}}$ needs to satisfy in Section 3.3 below. Finally, the adaptive LASSO-type estimator is given by

$$\hat{\mathbf{C}}^{aL} = \underset{\mathbf{\Psi}=\mathbf{\Psi}^\top, \mathbf{\Psi}>0}{\operatorname{argmin}} \; q_\lambda^{aL}(\mathbf{\Psi}). \tag{3.14}$$

[6] Note that other appropriate penalties could be chosen at this stage. However, as it will become clear the ℓ_1-penalty enables me to use efficient existing algorithms and it is not clear to what extend other penalties might improve the estimation results.

3.2.4 Choice of Pre-estimator for the Long-Run Covariance

As outlined in Sections 3.2.2 and 3.2.3, we are in dire need of an estimator for $\mathbf{\Sigma}$ in order to determine either the LASSO-type estimator (3.12) or the adaptive LASSO-type estimator (3.14). Throughout this chapter I make use of a particular HAC-estimator for long-run covariance matrices

$$\hat{\mathbf{\Sigma}} = T^{-1} \sum_{t=1}^{T} \sum_{s=1}^{T} k_\rho \left(\frac{|t-s|}{T} \right) \mathbf{y}_t \mathbf{y}_s^\top \tag{3.15}$$

where

$$k_\rho(x) = \begin{cases} (1-|x|)^\rho, & |x| \leq 1 \\ 0 & |x| > 1 \end{cases} \tag{3.16}$$

denotes the sharp origin kernel of Phillips et al. (2007). Note that in (3.15) the bandwidth for the kernel is set equal to the sample size rather than being, as usual, data dependent. This choice usually leads to inconsistent estimates for the long-run covariance matrix. However, in this particular case, the power parameter ρ in (3.16) takes the role of the bandwidth. As highlighted in Section 3.3, ρ is constructed such that it will grow along the sample size and the appropriate growth rates are given in the respective asymptotic results. Thus, as ρ grows, the sharp origin kernel $k_\rho(x)$ gives lower weight to higher order autocorrelations and, thereby, restores consistency of the estimator for $\mathbf{\Sigma}$, cf. Phillips et al. (2007) for more details. The reason why I choose the sharp origin kernel in (3.16) is that the proofs of the asymptotic results of the proposed LASSO-type estimators rely on the availability of an asymptotically normal pre-estimator $\hat{\mathbf{\Sigma}}$ for the long-run covariance $\mathbf{\Sigma}$ and, as far as I am aware, this is the only setting for which asymptotic normality of $\hat{\mathbf{\Sigma}}$ has been shown in the literature. Finally, proving such a statement for HAC-estimators with different kernels is outside of the scope of this study.

3.3 Asymptotic Properties

In order to be able to state asymptotic results about the proposed estimator some technical assumptions on the linear process $\mathbf{y}_t$, the sharp origin kernel $k_\rho(x)$ and the power parameter ρ need to be introduced.

On the linear process $\mathbf{y}_t$, I impose the same assumptions as Phillips et al. (2007). These are outlined below.

Assumption 3.1. (*i*) *The process* $\mathbf{y}_t$ *is zero-mean, fourth order stationary (i.e. its first, second, third and fourth moments are invariant to time shifts) and linear*

$$\mathbf{y}_t = \sum_{m=0}^{\infty} \mathbf{B}_m \boldsymbol{\epsilon}_{t-m}$$

with

$$\sum_{m=0}^{\infty} m^{1+\delta}\|\mathbf{B}_m\|_F < \infty, \quad \text{for some } \delta > 0.$$

where $\{\mathbf{B}_m \in \mathbb{R}^{N\times N}\}$ *is a sequence of parameter matrices.*

(ii) *The errors* $\boldsymbol{\epsilon}_t$ *satisfy*

$$\boldsymbol{\epsilon}_t \overset{i.i.d.}{\sim} (\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \quad \text{and} \quad \mathbb{E}\|\boldsymbol{\epsilon}_t\|_2^4 < \infty$$

(iii) *Moreover, the process* $\mathbf{y}_t$ *from* (i) *satisfies the functional central limit theorem*

$$T^{-1/2}\sum_{\tau=1}^{\lfloor Tr\rfloor} \mathbf{y}_\tau \Rightarrow \boldsymbol{\Lambda}\mathscr{B}_N(r), \quad r \in [0;1]$$

where $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top = \boldsymbol{\Sigma}$ *and* $\mathscr{B}_N(r)$ *is an* N*-dimensional vector of independent standard Brownian motions.*

On the power parameter ρ the following assumption, which is similar to the bandwidth expansion conditions found in the HAC-literature, is imposed.

Assumption 3.2. *The power parameter* ρ *satisfies*

$$\frac{1}{\rho} + \frac{\rho \ln T}{T} \to 0 \quad \text{as} \quad T \to \infty.$$

Assumption 1 is standard in the literature on HAC estimation. Part (i) restricts the parameter matrices of the linear process $\mathbf{y}_t$ and part (ii) specifies permissible innovations $\boldsymbol{\epsilon}_t$. In particular, part (ii) allows for conditional heteroskedastic linear processes. Together with Assumption 2, Phillips et al. (2007) verify their Theorem 3, which is stated below as Theorem 1 for completeness. This theorem is crucial to the results of this paper since the asymptotic results for the proposed LASSO-type estimators require that the estimator for $\boldsymbol{\Sigma}$ is asymptotically normal.

Theorem 3.1 (Phillips et al. (2007))**.** *Suppose that Assumptions 1 and 2 hold and that* $\rho = aT^b \to \infty$ *for some* $a > 0$ *and* $0 < b < \frac{2}{3}$*. Then*

$$\sqrt{\rho}(\operatorname{vec}\hat{\boldsymbol{\Sigma}} - \operatorname{vec}\boldsymbol{\Sigma}) \xrightarrow{D} \mathscr{N}(\mathbf{0}, (\mathbf{I}_{N^2} + \mathbf{K}_{N,N})(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}))$$

where $\mathbf{K}_{N,N}$ *is a* $N^2 \times N^2$*-commutation matrix that transforms* $\operatorname{vec}(\mathbf{W})$ *into* $\operatorname{vec}(\mathbf{W}^\top)$ *and* $\mathbf{I}_{N^2}$ *denotes the* $N^2 \times N^2$ *identity-matrix. That is,* $\mathbf{K}_{N,N} = \sum_{i=1}^N \sum_{j=1}^N \mathbf{e}_i\mathbf{e}_j^\top \otimes \mathbf{e}_j\mathbf{e}_i^\top$ *where* $\mathbf{e}_i$ *is the* $N \times 1$*-vector with* i*-th entry equal to one 1 and all other entries equal to zero, cf. Magnus and Neudecker (1979).*

Based on Assumptions 1 and 2, and Theorem 1 the asymptotic distribution of the LASSO-type estimator can be derived. Proposition 1 below states the result formally.

Theorem 3.2 (LASSO-type Estimator)**.** *Suppose Assumptions 1 and 2 hold, and that* $\rho = aT^b \to \infty$ *for some* $a > 0$ *and* $0 < b < \frac{2}{3}$*. If* $\sqrt{\rho}\lambda_T \to \lambda_0 \geq 0$ *as* $T \to \infty$*, the LASSO-type estimator defined in* (3.12) *satisfies*

$$\sqrt{\rho}(\hat{\mathbf{C}}^L - \mathbf{C}) \xrightarrow{D} \underset{\mathbf{U}=\mathbf{U}^\top}{\operatorname{argmin}} \Delta_q^L(\mathbf{U})$$

where

$$\begin{aligned}\Delta_q^L(\mathbf{U}) = {} & \operatorname{tr}(\mathbf{U}\boldsymbol{\Sigma}\mathbf{U}\boldsymbol{\Sigma}) + \operatorname{tr}(\mathbf{U}\mathbf{N}) \\ & + \lambda_0 \sum_{i \neq j} (u_{i,j} \operatorname{sgn}(c_{i,j}) \mathbb{1}\{c_{i,j} \neq 0\} + |u_{i,j}| \mathbb{1}\{c_{i,j} = 0\})\end{aligned}$$

and $\mathbf{N}$ *is a symmetric random* $N \times N$*-matrix such that* $\operatorname{vec}(\mathbf{N}) \sim \mathcal{N}(0, (\mathbf{I}_{N^2} + \mathbf{K}_{N,N})(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}))$

Thus, the convergence rate of the LASSO-type estimator is $T^{b/2}$, $0 < b < \frac{2}{3}$. Note that this is the same rate of convergence as for $\hat{\boldsymbol{\Sigma}}$ when the sharp origin kernel is used.

For the adaptive LASSO-type estimator, I derive the following result stated below.

Theorem 3.3 (Adaptive LASSO-type Estimator)**.** *Let* $\hat{\mathbf{C}}^{aL}$ *be as defined in* (3.14)*. Moreover, let* $\tilde{\mathbf{C}}$ *be a* $\sqrt{\rho}$*-consistent pre-estimator for* $\mathbf{C}$*,* $\rho\lambda_T \to \infty$*,* $\sqrt{\rho}\lambda_T \to 0$ *and* $\rho = aT^b \to \infty$ *for some* $a > 0$ *and* $0 < b < \frac{2}{3}$*. Then*

(i) $\hat{\mathbf{C}}^{aL}$ *consistently determines the index set of all zero entries in* $\mathbf{C}$*,* $\mathscr{S} = \{(i,j) : c_{i,j} = 0\}$*, i.e.*

$$\operatorname{Prob}(\hat{\mathscr{S}} = \mathscr{S}) \to 1$$

(ii) *the non-zero elements of* $\hat{\mathbf{C}}^{aL}$ *have the same limiting distribution as the oracle estimator,* $\hat{\mathbf{C}}^o$*, for which* $\mathscr{S}$ *is known. That is*

$$\sqrt{\rho}(\hat{\mathbf{C}}^o - \mathbf{C}) \xrightarrow{D} \underset{\mathbf{U}=\mathbf{U}^\top, u_{i,j}=0\,\forall\,(i,j)\in\mathscr{S}}{\operatorname{argmin}} \operatorname{tr}(\mathbf{U}\boldsymbol{\Sigma}\mathbf{U}\boldsymbol{\Sigma}) + \operatorname{tr}(\mathbf{U}\mathbf{N}).$$

In other words, Theorem 3.3 states that the adaptive LASSO-type estimator has the so-called oracle property of Zou (2006). That is, it identifies the sparsity structure $\mathscr{S}$ correctly with probability tending to one as the sample size grows, and provides an estimator for the non-zero elements of $\mathbf{C}$ with the same asymptotic distribution as the oracle estimator for which the true sparsity structure $\mathscr{S}$ is known. In light of statement (i) in Theorem 3.3, the use of the adaptive LASSO-type estimator is recommended when one is interested in recovering the sparsity structure of $\mathbf{C}$. This is further supported by the results from the Monte Carlo study in the next section.

3.4 Monte Carlo Simulation

In this section, an extensive Monte Carlo study is carried out to assess the small sample properties of the previously introduced estimators. The general simulation set up is as follows: The dimensions of the long-run precision matrices are such that $N \in \{10, 20\}$ is allowed and the sample sizes are set such that $T \in \{500, 1000, 2000, 3000, 4000, 5000\}$ with a separate burn-in-sample of 1000 observations, and, finally, 1000 Monte Carlo repetitions are carried out. Moreover, two classes of data generating processes (DGPs) are considered: a) vector autoregressive processes and b) vector moving average processes, since their long-run precision matrices have simple analytical solutions so that simulating data and assessing results obtained from these processes is straightforward. In addition, such processes are commonly used in empirical work. Finally, within each class two DGPs with specific long-run precision matrix structure are chosen. That is, in one case the long-run precision matrix is tridiagonal (representing a chain network) and in the second case its sparsity structure derives from the adjacency matrix of an Erdös-Rényi random graph. The latter is introduced in order to have a more realistic sparsity structure than the simple tridiagonal structure and is commonly used as a benchmark to evaluate the performance of precision matrix estimators. This allows to me to investigate whether the specific non-random tridiagonal structure affects the estimation results. More details are given below.

3.4.1 Data Generating Processes

In order to generate the long-run precision matrices, vector moving average (VMA) and autoregressive (VAR) processes are considered. This choice also allows to assess how different temporal dependence structures in the data affect the proposed estimator in small samples.

Vector Moving Average The first class of data generating processes (DGPs) considered is the class of vector moving average processes of order one:

$$\mathbf{y}_t = \boldsymbol{\epsilon}_t + \mathbf{B}\boldsymbol{\epsilon}_{t-1}, \quad \boldsymbol{\epsilon}_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_N) \tag{3.17}$$

where $\mathbf{B} \in \mathbb{R}^{N \times N}$ and the innovations' unconditional covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ is set to the $N \times N$ identity matrix for simplicity. The long-run covariance and precision matrices are then given by

$$\boldsymbol{\Sigma} = (\mathbf{I} + \mathbf{B})(\mathbf{I} + \mathbf{B})^\top \equiv \tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top \quad \text{and} \quad \mathbf{C} = (\tilde{\mathbf{B}}^{-1})^\top \tilde{\mathbf{B}}^{-1}. \tag{3.18}$$

Vector Autoregressive Process The second class of considered DGPs is a vector autoregression of order one:

$$\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_N) \tag{3.19}$$

where $\mathbf{A} \in \mathbb{R}^{N \times N}$ and the innovations' unconditional covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ is again set to $\mathbf{I}_N$ for simplicity. The long-run covariance and precision matrices are then given by

$$\boldsymbol{\Sigma} = [(\mathbf{I} - \mathbf{A})^\top (\mathbf{I} - \mathbf{A})]^{-1} \equiv [\tilde{\mathbf{A}}^\top \tilde{\mathbf{A}}]^{-1} \quad \text{and} \quad \mathbf{C} = \tilde{\mathbf{A}}^\top \tilde{\mathbf{A}} \tag{3.20}$$

From expressions (3.18) and (3.20), it can easily be seen that the structure of $\mathbf{C}$ can be controlled by the model parameters $\mathbf{A}$, respectively $\mathbf{B}$ and the next two paragraphs briefly outline how this is done for this Monte Carlo study.

Tridiagonal Precision Matrices In order to guarantee that $\mathbf{C}$ is tridiagonal, $\mathbf{\Sigma}$ is generated such that its (i,j)-th element is given by $\sigma_{ij} = \exp(-a|s_i - s_j|)$, where $a > 0$ and $s_1 < s_2 < ... < s_N$. In particular, I set $a = 0.75$, $s_i - s_{i-1} \overset{i.i.d}{\sim} U(0.5,1)$ for $i = 2,3,...,N$ and $s_1 \sim U(0.5,1)$. Next, according to (3.18), $\tilde{\mathbf{B}}$ can be computed by means of the Cholesky-decomposition of $\mathbf{\Sigma}$ and, afterwards, the model parameters are obtained as $\mathbf{B} = \tilde{\mathbf{B}} - \mathbf{I}$, in case of a VMA(1)-process.

For the VAR(1)-process the model parameters $\mathbf{A}$ are obtained from setting $\mathbf{A} = \mathbf{I} - \tilde{\mathbf{A}}$ where $\tilde{\mathbf{A}}$ is the solution to the Cholesky decomposition of $\mathbf{C}$, cf. Equation (3.20). Using the parameters obtained in such a way, data is then generated according to (3.17), respectively, (3.19).

Erdös-Rényi Precision Matrices The second class of precision matrices is derived from an Erdös-Rényi random graph. In particular, to keep comparability with the tridiagonal precision matrices, the number of edges is fixed to $3N - 2$ and edges are drawn uniformly randomly from the set of possible edges. In this way, the number of zero and non-zero elements in $\mathbf{C}$ is constant over all trials for a given spatial dimension N. Once the sparsity structure of $\mathbf{C}$ is determined, values to the non-zero entries are assigned based on the following rule: First, all off-diagonal non-zero elements are drawn from a uniform distribution on the interval $[0.3; 0.8]$. Half of these entries are then chosen at random and multiplied by -1. Afterwards, in order to guarantee that $\mathbf{C}$ (and, hence, $\mathbf{\Sigma}$) is positive definite the diagonal element in row i is set such that it equals the sum of the absolute values in row i plus 0.001. This makes $\mathbf{C}$ diagonally dominant and together with symmetry ensures positive definiteness, cf. Horn and Johnson (2013, Theorem 6.1.10).

Conditional Heteroskedasticity In addition to the four DGPs defined in the previous four paragraphs I also consider cases in which the DGPs feature conditional heteroskedasticity. In particular, I augment the above four DGPs such that their innovation terms follow an independent ARCH(1)-structure. That is, each coordinate's innovation term is an ARCH(1)-process and all these ARCH(1)-processes are independent of one another. In addition, I chose the ARCH-parameter such that the unconditional variance of the innovations equals 1 and the same long-run precision matrices as above are obtained

$$\epsilon_{i,t} = z_{i,t}\sigma_{\epsilon_i,t} \tag{3.21}$$

$$z_{i,t} \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \quad z_{i,t} \perp z_{j,t}\ \forall i \neq j \tag{3.22}$$

$$\sigma^2_{\epsilon_i,t} = 0.5 + 0.5\epsilon^2_{i,t-1} \tag{3.23}$$

Table 3.1 summarizes the different DGPs[7].

Table 3.1: Summary of the different DGPs

	DGP1	DGP2	DGP3	DGP4
class	VMA	VAR	VMA	VAR
lag length	1	1	1	1
structure $\mathbf{C}$	tridiagonal	tridiagonal	Erdös-Rényi	Erdös-Rényi
$\#\mathscr{S}^{\mathbb{C}}$	3N-2	3N-2	3N-2	3N-2
$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$	$\mathbf{I}_N$	$\mathbf{I}_N$	$\mathbf{I}_N$	$\mathbf{I}_N$
Heterosk.	No	No	No	No
T	500, 1000, 2000, 3000, 4000, 5000			
N	10, 20			
	DGP5	DGP6	DGP7	DGP8
class	VMA	VAR	VMA	VAR
lag length	1	1	1	1
structure $\mathbf{C}$	tridiagonal	tridiagonal	Erdös-Rényi	Erdös-Rényi
$\#\mathscr{S}^{\mathbb{C}}$	3N-2	3N-2	3N-2	3N-2
$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$	$\mathbf{I}_N$	$\mathbf{I}_N$	$\mathbf{I}_N$	$\mathbf{I}_N$
Heterosk.	Yes	Yes	Yes	Yes
T	500, 1000, 2000, 3000, 4000, 5000			
N	10, 20			

3.4.2 Choice of Auxiliary Quantities

This section briefly outlines how the tuning parameter λ and the pre-estimator for $\boldsymbol{\Sigma}$, respectively $\mathbf{C}$ in case of the adaptive LASSO are chosen.

Regularization Parameter λ The regularization parameter λ is usually unknown and, therefore, must be chosen from the data. Since the data is dependent and this dependence structure matters for the long-run covariance matrix, and therefore also for the long-run precision matrix, classic cross-validation is not feasible. A cross-validation like criterion such as the one proposed in Bickel and Levina (2008) adapted to a time series setting is computationally too demanding. Therefore, I opt to follow a different, commonly applied approach and choose the regularization parameter by minimizing an appropriate BIC criterion, generally given by

$$\text{BIC}(\lambda) = -\ln\det(\hat{\mathbf{C}}_\lambda) + \text{tr}(\hat{\mathbf{C}}_\lambda \hat{\boldsymbol{\Sigma}}) + \frac{\ln(T)}{T}\,\text{DoF}_\lambda \tag{3.24}$$

[7]Since there is no guarantee that the above procedures yield stationary processes, I checked separately that each simulated process is in fact stationary.

where DoF_λ denotes the degrees of freedom with which the likelihood function is penalized in the BIC-criterion. An inspection of the graphical LASSO algorithm yields that the correct degrees of freedom correction for the BIC is given by the non-sparsity:

$$\text{DoF}_\lambda = \sum_{i,j} \mathbb{1}\{(\hat{\mathbf{C}}_\lambda)_{ij} \neq 0\}. \tag{3.25}$$

That is, each non-zero parameter is estimated separately and, therefore, must be penalized individually.

Pre-Estimator for the Long-Run Covariance Matrix The pre-estimator for the long-run covariance is given in Equation (3.15) and the kernel of choice in Equation (3.16). As mentioned in Sections 3.2.4 and 3.3, the power parameter ρ tends to infinity as the sample size grows, thereby taking over the role of the bandwidth parameter in conventional HAC-estimation. Therefore, ρ preferably needs to be chosen in a data dependent manner. I will follow the suggestion of Phillips et al. (2007) and Newey and West (1994). That is, the optimal power parameter $\hat{\rho}$ minimizes the asymptotic mean square error of $\hat{\mathbf{\Sigma}}$. For more details I refer to Phillips et al. (2007).

Pre-Estimator for the Adaptive LASSO Weights As mentioned in Section 3.2.3, the adaptive LASSO-type estimator needs a $T^{b/2}$-consistent, $0 < b < \frac{2}{3}$, pre-estimator to construct the weighted penalty terms. In principle $\hat{\mathbf{\Sigma}}^{-1}$ would do the job, as can bee seen from Theorem 3.1. Instead of using this pre-estimator, however, I opt for using the LASSO-type estimator $\hat{\mathbf{C}}^L$ as pre-estimator which is also a valid choice by Theorem 3.2. The rationale behind this choice is that the LASSO-type estimator already penalizes some of the entries in $\hat{\mathbf{C}}$ and this is information which can be used to further improve the adaptive LASSO-type estimator.

Pre-Whitening Finally, while running the simulations, I found that the data dependent power parameter $\hat{\rho}$, chosen on the original series $\mathbf{y}_t$, does not provide a good estimate for the true power parameter ρ. For example, this way of choosing the power parameter translates into decreasing estimation accuracy as the sample sizes increases.[8] Pre-whitening the data, however, remedied this issue. Pre-whitening is done as suggested in Andrews and Monahan (1992). That is, first I fitted a VAR(1)-model to the original time series $\mathbf{y}_t$. After this has been done, I obtained the residuals and applied the long-run covariance estimator given in (3.15) and (3.16) to those residual series. Afterwards, the data is recolored to obtain the final pre-estimator $\hat{\mathbf{\Sigma}}$, see Andrews and Monahan (1992) for more details.

3.4.3 Computational Information

This section briefly summarizes how $\hat{\lambda}$ is determined and, afterwards, how knowledge about the DGP-class, respectively the sparsity structure is included in the simulations, for comparison with

[8] These results are not reported for brevity but available from the author upon request.

the methods proposed in this paper.

Solving the Graphical LASSO In order to solve problem (3.12), respectively (3.14) the graphical LASSO algorithm of Friedman et al. (2008) is used. As already mentioned in Section 4.2 the regularization parameter λ is unknown and, therefore, is determined as the minimizer of the BIC criterion (3.24) based on a grid of possible values for λ[9]. There is no advice available in the literature on how to determine this grid in practice. Moreover, Theorems 3.2 and 3.3 only specify that the optimal λ satisfies $aT^{b/2}\lambda_T \to 0$ and $aT^b\lambda_T \to \infty$, $a > 0$ and $0 < b < \frac{2}{3}$, which also does not provide much information on the choice of λ_T. Therefore, I propose the below described iterated search over a grid of possible values for λ.

In order to avoid spending too much computation time on parts of the grid which are far from the BIC solution for λ the first grid is set up relatively coarse to initially narrow down possible values for λ. Based on the optimal penalization parameter found on this initial grid, I construct a second grid to find the final, optimal solution for λ and, thus, for $\mathbf{C}$. This procedure can be summarized as follows:

1. Set the end points of the initial grid $\Lambda^{(1)}$ as $\Lambda^{(1)}_{\min} \leftarrow \min_{i,j}|\tilde{\sigma}_{i,j}|$, and $\Lambda^{(1)}_{\max} \leftarrow \max_{i,j}|\tilde{\sigma}_{i,j}|$ where $\tilde{\sigma}_{i,j}$ denotes the (i,j)-th entry of $\hat{\mathbf{\Sigma}}^{-1}$.
2. Set $\Lambda^{(1)} \leftarrow \ln(\text{seq}(\text{from} = \exp(\Lambda^{(1)}_{\min}), \text{to} = \exp(\Lambda^{(1)}_{\max}), \text{length} = 100))$ where $\text{seq}(s, e, \#\lambda)$ generates a sequence from s to e with $\#\lambda$ points.
3. Set $\hat{\lambda}^{(1)} = \Lambda^{(1)}_{k^*} \leftarrow \text{argmin}_\lambda \text{BIC}(\lambda)$ where k^* denotes the position of $\hat{\lambda}^{(1)}$ in $\Lambda^{(1)}$
4. Set $\hat{\lambda}^{(1)}_{-1} \leftarrow \Lambda^{(1)}_{k^*-1}$ and $\hat{\lambda}^{(1)}_{+1} \leftarrow \Lambda^{(1)}_{k^*+1}$
5. Update the end points of the grid as $\Lambda^{(2)}_{\min} \leftarrow \hat{\lambda}^{(1)}_{-1}$ and $\Lambda^{(2)}_{\max} \leftarrow \hat{\lambda}^{(1)}_{+1}$
6. Repeat 2 and 3 to find the final $\hat{\lambda}^{(2)}$.

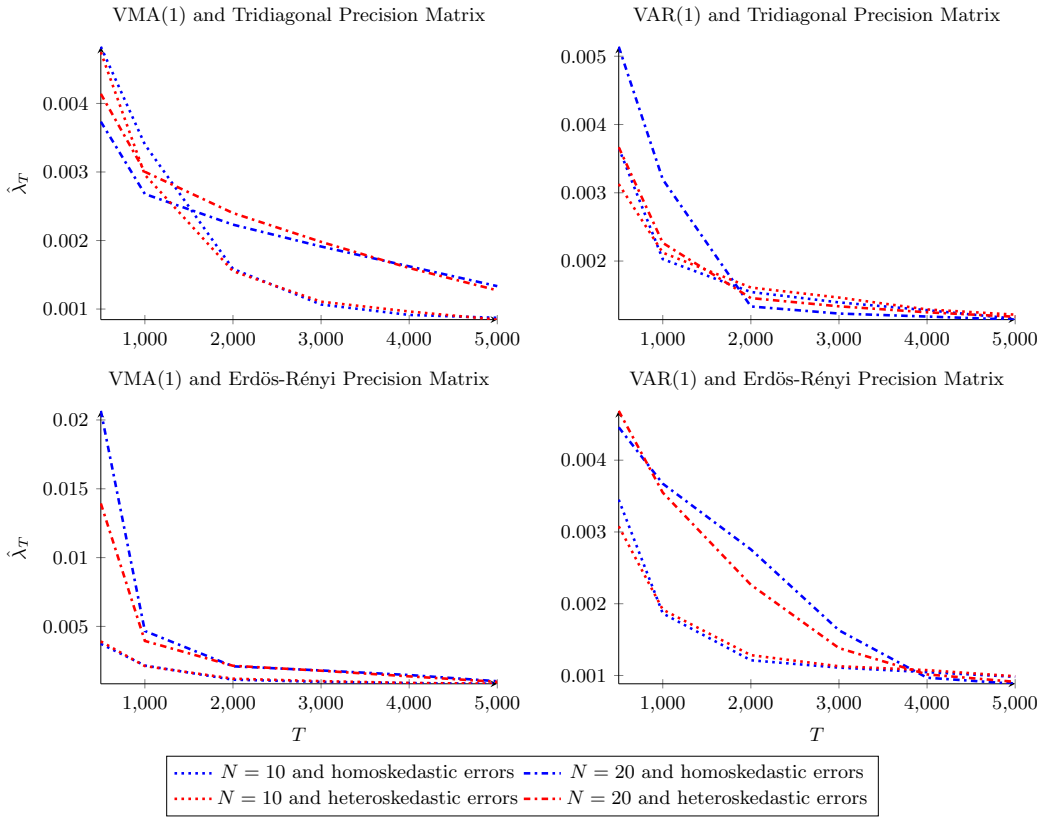
The idea behind this approach is as follows: If there exists an unique minimizer of the BIC in (3.24)[10] then the procedure must find a point on the first grid that is closest to the minimizer in the sense of having a small associated BIC value. Since the minimizer is usually assumed to be unique it then also follows, that it must lie in the interval spanned by the two neighboring points of this initially found grid-point. Thus, a second step is applied to further narrow down the true minimizer of the BIC.

In Figure 3.3 below I plot the Monte Carlo averages of the penalization parameter λ obtained by the above procedure. These plots illustrate that the values of λ are in line with the

[9] The choice of λ crucially affects the results, and more research is needed to uncover the best choice of λ.

[10] Note that the LASSO literature rarely talks about the existence of an unique BIC minimizing value for λ which satisfies the theoretical requirements. The BIC, or any other criterion for this matter, is usually applied because it works well in simulations.

Figure 3.3: Monte Carlo Means of $\hat{\lambda}_T$ for the adaptive LASSO



requirements of Theorem 3.3. That is, for all 8 DGPs $\hat{\lambda}_T$ decreases as the sample size increases so that $\sqrt{\hat{\rho}}\hat{\lambda}_T \to \lambda_0 \geq 0$ and $\hat{\rho}\hat{\lambda}_T \to \infty$ are possible.

Vector Moving Average In case of a VMA(q)-process, knowledge about the DGP can be incorporated into the estimation procedure via the pre-estimator for the long-run covariance matrix $\hat{\mathbf{\Sigma}}$ since it holds that

$$\mathbf{\Sigma} = \sum_{h=-q}^{q} \mathbf{\Gamma}(h). \tag{3.26}$$

Thus, a VMA(1)-DGP implies that the long-run covariance matrix consists of only the short-run covariance and the autocovariance at lag order one. Thus, rather than using the HAC pre-estimator, the pre-estimator can directly be based on a simplified expression for the long-run covariance matrix

$$\hat{\mathbf{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{y}_t\mathbf{y}_t^\top + \frac{1}{2T}\sum_{t=1}^{T-1} \mathbf{y}_{t+1}\mathbf{y}_t^\top + \left(\frac{1}{2T}\sum_{t=1}^{T-1} \mathbf{y}_{t+1}\mathbf{y}_t^\top\right)^\top \tag{3.27}$$

where the factor 0.5 in the last two summands is introduced to guarantee that the resulting estimator is positive definite[11].

[11] So, it coincides with the Bartlett kernel with known lag length $q = 1$

Vector Autoregression In case of a VAR(1)-process, knowledge about the the DGP can be incorporated by recalling that the long-run precision matrix is given by

$$\hat{\mathbf{C}} = \left(\mathbf{I}_N - \hat{\mathbf{A}}\right)^\top \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \left(\mathbf{I}_N - \hat{\mathbf{A}}\right). \tag{3.28}$$

That is, the long-run precision matrix can be estimated by estimating $\mathbf{A}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}$, for which reliable methods exists. Either, one can estimate $\mathbf{A}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}$ separately by first applying an adaptive LASSO equation-by-equation to the VAR(1)-process to obtain $\hat{\mathbf{A}}$, obtaining the residuals and then applying, e.g. the graphical LASSO on the residuals to obtain $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}$. Moreover, Barigozzi and Brownlees (2017) propose to estimate $\mathbf{A}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}$ jointly in their `nets`-algorithm. For the Monte Carlo study I opt for the first approach since the `nets`-algorithm involves a computationally costly two-dimensional grid search and it is not clear how performance is improved by estimating the quantities of interest jointly. Finally, one should note that it is not entirely clear a priori whether this procedure indeed improves the resulting estimator for $\mathbf{C}$ over the proposed method since estimation errors (especially in the sparsity structure) in both $\hat{\mathbf{A}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}$ can contaminate the final estimator due to the multiplicative structure in (3.28) and, thereby, could worsen the results.

Estimation of the Long-Run Precision Matrix when the Sparsity Pattern is Known In this case the sparsity pattern is known, but not the DGP-class. This information can be incorporated in the proposed estimation scheme in the following way. First, the pre-estimator which is parsed to the objective functions (3.11) and (3.13) is the above HAC estimator with the sharp origin kernel. Since the sparsity pattern is known, that is the index set $\mathscr{S}$ is known, a differential penalty can be applied. That is, for all pairs of indices $(i^*, j^*) \in \mathscr{S}$ a penalty of infinity is applied, thereby forcing the associated estimated values $\hat{c}_{i^*,j^*}$ to be zero. On the other hand for all indices $(i^*, j^*) \in \mathscr{S}^\complement$ the penalty parameter is set to 0 and thus, none of these entries is penalized.

3.4.4 Results

The performance of the proposed estimator is evaluated according to three criteria. First, the Frobenius-norm of the difference between the estimator and the true long-run precision matrix is computed, thereby measuring the overall closeness of the estimator and the true quantity of interest. Second, the Type I error rate (defined as the number of entries in the long-run precision matrix which are estimated to be zero but are non-zero in reality divided by the true amount of true non-zero elements), committed during estimation, is computed. Third, the Type II error rates (defined as the number of entries in the precision matrix which are estimated to be non-zero but are zero in reality divided by the amount of true zero entries) is determined. The latter two criteria allow assessment of the covariance selection, that is the capability of differentiating between true zero and non-zero entries, properties of the proposed estimators. Ideally, the Type I

and II error rates are as close to zero as possible.

Moreover, in order to assess how having prior information about either the DGP-class or the sparsity structure of the true long-run precision matrix affects the estimators, the following three scenarios are considered: Firstly, the long-run precision matrix is estimated by being completely agnostic about both the underlying DGP-class and the sparsity structure. Thus, this scenario constitutes the most common case empirical researchers tackle. The second scenario assumes that the underlying DGP-class is known (i.e. whether the data is generated by either a VMA(1)- or a VAR(1)-process) but not the sparsity structure of the long-run precision matrix. Lastly, the third scenario assumes that the sparsity structure is known, but not the underlying DGP-class (obviously, for this case no Type I and II errors are reported). While discussing the results, I will refer to the latter case as the oracle estimator since $\mathscr{S}$ is known.

Below, the results of this Monte Carlo study are discussed. While doing so, I always combine the two DGPs which are in the same column in Table 3.1. That is, both considered DGPs only vary in one aspect, namely whether their error terms are conditionally homo- or heteroskedastic. The results are reported in Appendix B. The tables display averages of the respective quantities over 1000 Monte Carlo repetitions. Standard errors over these repetitions are presented in parentheses.

DGP1 and DGP5 The results for these two VMA(1) process with a tridiagonal precision matrix can be found in Tables B.1–B.3. In terms of the norm difference it can be seen that it decreases in all cases when the sample size increases. Moreover, the oracle estimator performs best with lowest norm differences overall. Plainly inverting $\hat{\mathbf{\Sigma}}$ performs as well as both LASSO-type estimators and knowing the DGP-class performs worst. Moreover, the norm differences are larger when the dimension of $\mathbf{C}$ is larger. However, this comes at no surprise since more parameters need to be estimated.

In terms of Type I error rates we observe that all of them are virtually equal to zero, except for the case where it is known that the data follows a VMA(1)-process. However, even in this case they quickly converge to (almost) zero.

In terms of Type II error rates the most striking result is probably that they increase with the sample size when the DGP-class is known. The proposed LASSO-type estimators behave differently. In particular, the plain LASSO estimator provides almost no penalization with a stable Type II error rate of about 95%. The adaptive LASSO estimator performs much better but still allows for a Type II error rate of about 40% for $T = 500$ which decreases as the sample size increases. However, another possible explanation is that in the case of a known VMA(1) DGP-class the LASSO-type estimators tend to underpenalize.

Finally, there are no essential differences between the conditionally homoskedastic and heteroskedastic cases.

DGP2 and DGP6 In case of a VAR(1)-process with tridiagonal long-run precision matrix it can again be observed that the norm differences decrease as the sample size increases, no matter which dimenson of $\mathbf{C}$ or types of error terms are considered, see Table B.4. In this case, however, when one knows the VAR(1) DGP-class one obtains the smallest error, followed by the infeasible oracle estimator. Finally, the adaptive LASSO-type estimator ranks third, followed by the LASSO and the inverse long-run covariance estimator. Note that the differences between the leading three estimators become less pronounced when the data is conditionally heteroskedastic.

In terms of Type I error rates we again observe that they all are virtually zero in all cases, see Table B.5. The Type II error rates, in contrast, can be quite substantial, see Table B.6. In particular, the LASSO-type estimator performs worst with error rates of about 90% and they seem to be non-decreasing in the sample size. The adaptive LASSO-type estimator is the second best with error rates decreasing to about 12% for $T = 5000$. Finally, the parametric estimator provides the lowest error rates which decrease from about 5% to almost 0% as the sample size increases.

Again, differences between the conditionally homo- and heteroskedasticity cases are relatively small. Finally, the results in this case can again be explained by underpenalization during the proposed estimation procedures.

DGP3 and DGP7 The results for the VMA(1)-case can be found in Tables B.7–B.9. The patterns do not change much compared to the VMA(1) case with a tridiagonal long-run covariance matrix. As before, the norm differences decrease as the sample size increases and the oracle estimator performs best, followed by the two LASSO-type estimators and inverse long-run sample covariance. The case where the VMA(1) DGP-class is known comes in last.

For the Type I error rates we observe the already familiar pattern. For all estimators this criterion equals virtually zero. The Type II error rates are again relatively large, but still acceptable for the adaptive LASSO-type estimator, especially in light of the apparent underpenalization. Note that for this estimator the error rate stays stable over the sample size. For the LASSO-type estimators this result can also be explained by how the data is pre-whitened. As mentioned in Section 4.2 I opted to pre-whiten by using a VAR(1) approximation. However, the fact that a VMA(1)-process has a VAR(∞)-representation suggests that a higher order VAR-process in the pre-whitening stage can yield better results in this case. Finally, it does not seem to matter for the results whether the errors are homo- or heteroskedastic.

DGP4 and DGP8 The last two cases are a VAR(1)-process with an Erdös-Rényi type long-run precision matrix. The results can be found in Tables B.10–B.12. Again, the norm differences decrease as the sample size increases. The best performing estimator is again the case where the DGP-class is known, followed by the oracle and adaptive LASSO-type estimator. Again, in the case of conditionally heteroskedastic errors the adaptive LASSO-type estimator is loser to the two leading estimators with respect to norm differences.

Regarding the Type I error rates, they are again virtually equal to zero for all cases. The Type II error rates also follow the same pattern as for DGP2 and DGP6. That is, they decrease with the sample size increasing and are lowest for the estimator for which the DGP-class is known, followed by the adaptive LASSO-type estimator.

Summarizing the results of this Monte Carlo study, it seems that it does not matter for the proposed estimators whether the process is conditionally homo- or heteroskedastic. In addition, the adaptive LASSO-type estimator performs well compared to the naïve estimator $\hat{\mathbf{\Sigma}}^{-1}$ in cases where there is no knowledge about the DGP or the sparsity structure. In case where it is known that the true DGP is a VAR-process, it seems that a parametric estimator is favorable over the proposed non-parametric estimators. However, the same cannot be said when the DGP is in the VMA-class. Thus, one intends to use the highly recommended pre-whitening step, it could favorable to determine the VAR-order in a data dependent way, for example via the BIC, to be more robust against possible Type II errors. Finally, the tendency to underpenalize can be tackled by examining more closely the choice of $\hat{\lambda}$ through additional simulations.

3.5 Conclusion

In this chapter I propose a novel estimator for sparse long-run precision matrices for possibly conditionally heteroskedastic linear time series. Under standard assumptions I show that both LASSO-type estimator are $T^{b/2}$-consistent with $0 < b < \frac{2}{3}$, and that the adaptive LASSO-type estimator has the oracle property of Zou (2006), where the convergence rate of $T^{b/2}$ derives from the choice of the sharp origin kernel for the pre-estimator of the long-run covariance of the time series. Finally, I assess the small sample performance of the proposed estimator by means of an extensive Monte Carlo study and find that it performs fairly well in small samples with a tendency to underpenalize, but almost never sets elements to zero that are non-zero in reality.

For future research, extensions to other, commonly used kernels, such as the Quadratic Spectral kernel might be of interest since they are expected to provide faster convergence rates. Moreover, different penalties, such as SCAD and MCP, or different Bregman-divergences are also a valuable choice to further improve the proposed estimator. Finally, given the ever growing availability of data, an extension to a high-dimensional setting where N is allowed to grow might also be of future interest.

Appendix 3.A Mathematical Proofs

The proofs of Propositions 1 and 2 closely follow those of Yuan and Lin (2007). These authors proof a similar result to that in Proposition 1 for the LASSO-type estimator and results similar to

that of Proposition 2 for a non-negative Garrote-type estimator[12] based on $i.i.d.$ normal random vectors.

PROOF OF THEOREM 3.2. Let $\mathbf{\Psi} = \mathbf{C} + \frac{\mathbf{U}}{\sqrt{\rho}}$ where $\mathbf{U} = \mathbf{U}^\top$ and define $\Delta^L_{q,T}(\mathbf{U}) = q^L_\lambda(\mathbf{\Psi}) - q^L_\lambda(\mathbf{C})$. Then,

$$\begin{aligned}\Delta^L_{q,T}(\mathbf{U}) = &-\log\left|\mathbf{C} + \frac{\mathbf{U}}{\sqrt{\rho}}\right| + \operatorname{tr}\left[\left(\mathbf{C} + \frac{\mathbf{U}}{\sqrt{\rho}}\right)\hat{\mathbf{\Sigma}}\right] + \lambda_T \sum_{i\neq j}\left|\frac{u_{i,j}}{\sqrt{\rho}} + c_{i,j}\right| \\ &+ \log|\mathbf{C}| - \operatorname{tr}(\mathbf{C}\hat{\mathbf{\Sigma}}) - \lambda_T \sum_{i\neq j}|c_{i,j}|. \end{aligned} \tag{3.29}$$

Let $\mu_i(\mathbf{M})$ denote the i-th largest eigenvalue of a symmetric matrix $\mathbf{M}$ and notice that

$$\begin{aligned}\log\left|\mathbf{C} + \frac{\mathbf{U}}{\sqrt{\rho}}\right| - \log|\mathbf{C}| &= \log\left|\mathbf{I}_N + \frac{\mathbf{\Sigma}^{1/2}\mathbf{U}\mathbf{\Sigma}^{1/2}}{\sqrt{\rho}}\right| \\ &= \sum_{i=1}^N (1 + \mu_i(\mathbf{\Sigma}^{1/2}\mathbf{U}\mathbf{\Sigma}^{1/2})/\sqrt{\rho}) \\ &= \sum_{i=1}^N \frac{\mu_i(\mathbf{\Sigma}^{1/2}\mathbf{U}\mathbf{\Sigma}^{1/2})}{\sqrt{\rho}} - \frac{\mu_i^2(\mathbf{\Sigma}^{1/2}\mathbf{U}\mathbf{\Sigma}^{1/2})}{\rho} + o(\rho^{-1}) \\ &= \frac{\operatorname{tr}(\mathbf{\Sigma}^{1/2}\mathbf{U}\mathbf{\Sigma}^{1/2})}{\sqrt{\rho}} - \frac{\operatorname{tr}(\mathbf{\Sigma}^{1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{U}\mathbf{\Sigma}^{1/2})}{\rho} + o(\rho^{-1}) \\ &= \frac{\operatorname{tr}(\mathbf{U}\mathbf{\Sigma})}{\sqrt{\rho}} - \frac{\operatorname{tr}(\mathbf{U}\mathbf{\Sigma}\mathbf{U}\mathbf{\Sigma})}{\rho} + o(\rho^{-1}) \end{aligned} \tag{3.30}$$

and that

$$\operatorname{tr}\left[\left(\mathbf{C} + \frac{\mathbf{U}}{\sqrt{\rho}}\right)\hat{\mathbf{\Sigma}}\right] - \operatorname{tr}(\mathbf{C}\hat{\mathbf{\Sigma}}) = \operatorname{tr}\left(\frac{\mathbf{U}\mathbf{\Sigma}}{\sqrt{\rho}}\right) + \operatorname{tr}\left(\frac{\mathbf{U}(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})}{\sqrt{\rho}}\right). \tag{3.31}$$

Moreover, for sufficiently large $\sqrt{\rho}$ it holds that

$$\begin{aligned}&\lambda_T \sum_{i\neq j}\left|\frac{u_{i,j}}{\sqrt{\rho}} + c_{i,j}\right| - \lambda_T \sum_{i\neq j}|c_{i,j}| \\ &= \frac{\lambda_T}{\sqrt{\rho}} \sum_{i\neq j}\Big(u_{i,j}\operatorname{sgn}(c_{i,j})\mathbb{1}\{c_{i,j}\neq 0\} + |u_{i,j}|\mathbb{1}\{c_{i,j} = 0\}\Big). \end{aligned} \tag{3.32}$$

Combining (3.30)–(3.32) with (3.29) directly yields

$$\begin{aligned}\rho\Delta^L_{q,T}(\mathbf{U}) = &\operatorname{tr}(\mathbf{U}\mathbf{\Sigma}\mathbf{U}\mathbf{\Sigma}) + \operatorname{tr}\Big(\mathbf{U}\sqrt{\rho}(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\Big) + o(1) \\ &+ \sqrt{\rho}\lambda_T \sum_{i\neq j}\Big(u_{i,j}\operatorname{sgn}(c_{i,j})\mathbb{1}\{c_{i,j}\neq 0\} + |u_{i,j}|\mathbb{1}\{c_{i,j} = 0\}\Big). \end{aligned} \tag{3.33}$$

[12] The difference between the adaptive LASSO-type and the non-negative Garrote-type estimator lies in the penalty term. In particular, the non-negative Garrote-type penalty is given by $\lambda_T \sum_{i\neq j}\psi_{i,j}/\tilde{c}_{i,j}$ subject to $\psi_{i,j}/\tilde{c}_{i,j} \geq 0$ whereas the adaptive LASSO-type penalty is given by $\lambda_T \sum_{i\neq j}|\psi_{i,j}|/|\tilde{c}_{i,j}|$.

By Theorem 1, $\sqrt{\rho}(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})$ converges in distribution to a multivariate zero-mean normal random matrix $\mathbf{N}$ with covariance matrix as provided in Theorem 1. Moreover, $\lambda_T\sqrt{\rho} \to \lambda_0 \geq 0$ by assumption. Therefore,

$$\begin{aligned}\rho\Delta_{q,T}^{L}(\mathbf{U}) &\xrightarrow{D} \operatorname{tr}(\mathbf{U\Sigma U\Sigma}) + \operatorname{tr}\Big(\mathbf{UN}\Big) \\ &\quad + \lambda_0 \sum_{i\neq j}\Big(u_{i,j}\operatorname{sgn}(c_{i,j})\mathbb{1}\{c_{i,j}\neq 0\} + |u_{i,j}|\mathbb{1}\{c_{i,j}=0\}\Big) \\ &\equiv \Delta_q^L(\mathbf{U}). \end{aligned} \tag{3.34}$$

Finally, $\rho\Delta_{q,T}^{L}(\mathbf{U})$ and $\Delta_q^L(\mathbf{U})$ in (3.34) are both convex functions and the latter has a unique minimum.[13] Therefore, it follows that

$$\operatorname*{argmin}_{\mathbf{U}=\mathbf{U}^\top} \rho\Delta_{q,T}(\mathbf{U}) = \sqrt{\rho}(\hat{\mathbf{C}} - \mathbf{C}) \xrightarrow{D} \operatorname*{argmin}_{\mathbf{U}=\mathbf{U}^\top} \Delta_q^L(\mathbf{U}) \tag{3.35}$$

concluding the proof. □

PROOF OF THEOREM 3.3. As in the Proof of Theorem 3.2, one can show that

$$\begin{aligned}\rho\Delta_{q,T}^{aL}(\mathbf{U}) &= \operatorname{tr}(\mathbf{U\Sigma U\Sigma}) + \operatorname{tr}\Big(\mathbf{U}\sqrt{\rho}(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\Big) + o(1) \\ &\quad + \sqrt{\rho}\lambda_T \sum_{i\neq j} \sqrt{\rho}\left(\left|c_{i,j} + \frac{u_{i,j}}{\sqrt{\rho}}\right| - |c_{i,j}|\right)/|\tilde{c}_{i,j}|. \end{aligned} \tag{3.36}$$

Now, there are two cases to consider: *a*) $c_{i,j} \neq 0$ and *b*) $c_{i,j} = 0$. In case *a*) it holds that $|\tilde{c}_{i,j}|^{-1} \xrightarrow{p} |c_{i,j}|^{-1}$, $\sqrt{\rho}\Big(|c_{i,j} + u_{i,j}/\sqrt{\rho}| - |c_{i,j}|\Big) \to u_{i,j}\operatorname{sgn}(c_{i,j})$ and $\sqrt{\rho}\lambda_T \to 0$ by assumption. So, by Slutksy's Theorem the penalty term in (3.36) converges to 0 for all $(i,j) \in \mathscr{S}^{\complement}$.
On the other hand, in case *b*) it holds that $\sqrt{\rho}\Big(|c_{i,j} + u_{i,j}/\sqrt{\rho}| - |c_{i,j}|\Big) = |u_{i,j}|$ and $\sqrt{\rho}\lambda_T|\tilde{c}_{i,j}|^{-1} = \frac{\rho\lambda_T}{|\sqrt{\rho}\tilde{c}_{i,j}|}$ where $\sqrt{\rho}\tilde{c}_{i,j} = \mathscr{O}_p(1)$. Therefore, Equation (3.36) can be rewritten as

$$\rho\Delta_{q,T}^{aL}(\mathbf{U}) = \operatorname{tr}(\mathbf{U\Sigma U\Sigma}) + \operatorname{tr}\Big(\mathbf{U}\sqrt{\rho}(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\Big) + \rho\lambda_T \sum_{(i,j)\in\mathscr{S}} \frac{|u_{i,j}|}{|\sqrt{\rho}\tilde{c}_{i,j}|} + o(1). \tag{3.37}$$

Since $\rho\lambda_T \to \infty$ by Assumption, it must be true that the minimizer of $\rho\Delta_{q,T}^{aL}(\mathbf{U})$ is such that $u_{i,j} = 0$ whenever $(i,j) \in \mathscr{S}$ with probability tending to one. Otherwise $\rho\Delta_{q,T}(\mathbf{U})$ would diverge to infinity and this is in contradiction with Theorem 3.2.

The limiting distribution of $\hat{\mathbf{C}}^{aL}$ is easily dervied by noting that the pseudo ML estimator, $\hat{\mathbf{C}}^{o}$, based on the true sparsity structure $\mathscr{C}$ is such that

$$\sqrt{\rho}(\hat{\mathbf{C}}^{o} - \mathbf{C}) \xrightarrow{D} \operatorname*{argmin}_{\mathbf{U}=\mathbf{U}^\top,\, u_{i,j}=0\,\forall\in\mathscr{S}} \operatorname{tr}(\mathbf{U\Sigma U\Sigma}) + \operatorname{tr}\Big(\mathbf{UN}\Big). \tag{3.38}$$

□

[13] Note that $\operatorname{tr}(\mathbf{U\Sigma U\Sigma})$ is a term which is quadratic in $\mathbf{U}$ and $\operatorname{tr}(\mathbf{UN})$ is linear in $\mathbf{U}$. Thus, (3.34) constitute a parabola defined on the space of all symmetric matrices $\mathbf{U}$.

Appendix 3.B Tables

Table 3.2: VMA(1) with Tridiagonal Precision Matrix – Norm Differences

		$T = 500$	$T = 1000$	$T = 2000$	$T = 3000$	$T = 4000$	$T = 5000$
Conditional Homoskedastic Errors							
$N = 10$	$\hat{\mathbf{\Sigma}}^{-1}$	2.42	2.33	2.23	2.16	2.11	2.06
		(0.16)	(0.12)	(0.09)	(0.08)	(0.07)	(0.07)
	$\hat{\mathbf{C}}^{L}$	2.43	2.35	2.24	2.17	2.11	2.07
		(0.16)	(0.12)	(0.10)	(0.08)	(0.07)	(0.07)
	$\hat{\mathbf{C}}^{aL}$	2.44	2.34	2.24	2.16	2.11	2.06
		(0.18)	(0.13)	(0.10)	(0.08)	(0.07)	(0.07)
	$\hat{\mathbf{C}}^{o}$	1.77	1.72	1.64	1.58	1.53	1.49
		(0.19)	(0.14)	(0.11)	(0.09)	(0.08)	(0.07)
	$\hat{\mathbf{C}}^{C}$	4.53	4.36	3.86	3.5	3.36	3.27
		(0.10)	(0.19)	(0.28)	(0.17)	(0.09)	(0.07)
$N = 20$	$\hat{\mathbf{\Sigma}}^{-1}$	4.60	4.40	4.20	4.07	3.96	3.87
		(0.16)	(0.12)	(0.10)	(0.08)	(0.08)	(0.07)
	$\hat{\mathbf{C}}^{L}$	4.60	4.42	4.22	4.09	3.99	3.90
		(0.17)	(0.12)	(0.10)	(0.08)	(0.08)	(0.07)
	$\hat{\mathbf{C}}^{aL}$	4.57	4.41	4.22	4.09	3.98	3.89
		(0.22)	(0.13)	(0.11)	(0.09)	(0.08)	(0.07)
	$\hat{\mathbf{C}}^{o}$	3.52	3.41	3.23	3.11	3.01	2.92
		(0.21)	(0.15)	(0.12)	(0.09)	(0.09)	(0.08)
	$\hat{\mathbf{C}}^{C}$	7.79	7.54	6.77	6.23	5.65	5.48
		(0.07)	(0.18)	(0.15)	(0.41)	(0.15)	(0.08)
Conditional Heteroskedastic Errors							
$N = 10$	$\hat{\mathbf{\Sigma}}^{-1}$	2.45	2.33	2.23	2.16	2.10	2.05
		(0.22)	(0.19)	(0.15)	(0.14)	(0.12)	(0.11)
	$\hat{\mathbf{C}}^{L}$	2.46	2.34	2.24	2.16	2.11	2.06
		(0.22)	(0.19)	(0.15)	(0.14)	(0.12)	(0.11)
	$\hat{\mathbf{C}}^{aL}$	2.46	2.34	2.23	2.16	2.10	2.05
		(0.26)	(0.2)	(0.15)	(0.14)	(0.12)	(0.12)
	$\hat{\mathbf{C}}^{o}$	1.78	1.71	1.63	1.57	1.52	1.48
		(0.31)	(0.25)	(0.19)	(0.17)	(0.15)	(0.14)
	$\hat{\mathbf{C}}^{C}$	4.46	4.23	3.75	3.46	3.33	3.24
		(0.20)	(0.30)	(0.32)	(0.19)	(0.13)	(0.12)
$N = 20$	$\hat{\mathbf{\Sigma}}^{-1}$	4.62	4.38	4.18	4.05	3.94	3.86
		(0.22)	(0.19)	(0.15)	(0.14)	(0.13)	(0.13)
	$\hat{\mathbf{C}}^{L}$	4.60	4.39	4.21	4.07	3.97	3.88
		(0.23)	(0.19)	(0.15)	(0.14)	(0.13)	(0.13)
	$\hat{\mathbf{C}}^{aL}$	4.55	4.38	4.20	4.07	3.96	3.88
		(0.26)	(0.2)	(0.16)	(0.15)	(0.13)	(0.13)

Continued on next page

Table 3.2 – continued from previous page

		$T = 500$	$T = 1000$	$T = 2000$	$T = 3000$	$T = 4000$	$T = 5000$
	$\hat{\mathbf{C}}^o$	3.47	3.36	3.21	3.09	2.99	2.91
		(0.33)	(0.26)	(0.19)	(0.18)	(0.16)	(0.15)
	$\hat{\mathbf{C}}^C$	7.70	7.38	6.66	5.97	5.60	5.45
		(0.16)	(0.27)	(0.35)	(0.41)	(0.16)	(0.14)

[1] Monte Carlo standard errors are presented in parantheses below the Monte Carlo means of the criterion.

[2] $\hat{\mathbf{\Sigma}}^{-1}$ denotes the inverse of the long-run covariance estimator, $\hat{\mathbf{C}}^L$ the LASSO, $\hat{\mathbf{C}}^{aL}$ the adaptive LASSO, $\hat{\mathbf{C}}^o$ the oracle and $\hat{\mathbf{C}}^C$ the known DGP-class estimator

Table 3.3: VMA(1) with Tridiagonal Precision Matrix – Type 1 Error Rates

		$T = 500$	$T = 1000$	$T = 2000$	$T = 3000$	$T = 4000$	$T = 5000$
Conditional Homoskedastic Errors							
$N = 10$	$\hat{\mathbf{C}}^L$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^{aL}$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^C$	0.43	0.23	0.02	0.00	0.00	0.00
		(0.09)	(0.17)	(0.04)	(0.01)	(0.00)	(0.00)
$N = 20$	$\hat{\mathbf{C}}^L$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^{aL}$	0.00	0.00	0.00	0	0.00	0.00
		(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^C$	0.45	0.23	0.02	0.01	0.00	0.00
		(0.05)	(0.11)	(0.02)	(0.01)	(0.00)	(0.00)
Conditional Heteroskedastic Errors							
$N = 10$	$\hat{\mathbf{C}}^L$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^{aL}$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.02)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^C$	0.38	0.16	0.02	0.00	0.00	0.00
		(0.13)	(0.15)	(0.03)	(0.01)	(0.00)	(0.00)
$N = 20$	$\hat{\mathbf{C}}^L$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^{aL}$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^C$	0.41	0.17	0.02	0.00	0.00	0.00
		(0.08)	(0.11)	(0.02)	(0.01)	(0.00)	(0.00)

[1] Monte Carlo standard errors are presented in parantheses below the Monte Carlo means of the criterion.

[2] $\hat{\mathbf{C}}^L$ denotes the LASSO, $\hat{\mathbf{C}}^{aL}$ the adaptive LASSO and $\hat{\mathbf{C}}^C$ the known DGP-class estimator when it is known that the data is generated by a VMA(1)-process.

[3] This table reports the error rates as fractions between committed Type I errors and the amount of true entries unequal to zero on the lower triangle of the matrix.

Table 3.4: VMA(1) with Tridiagonal Precision Matrix – Type 2 Error Rates

		$T = 500$	$T = 1000$	$T = 2000$	$T = 3000$	$T = 4000$	$T = 5000$
Conditional Homoskedastic Errors							
$N = 10$	$\hat{\mathbf{C}}^L$	0.92	0.93	0.95	0.96	0.96	0.96
		(0.05)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)
	$\hat{\mathbf{C}}^{aL}$	0.46	0.46	0.52	0.54	0.53	0.52
		(0.11)	(0.10)	(0.10)	(0.08)	(0.07)	(0.07)
	$\hat{\mathbf{C}}^C$	0.00	0.02	0.15	0.25	0.26	0.26
		(0.02)	(0.07)	(0.16)	(0.10)	(0.07)	(0.07)
$N = 20$	$\hat{\mathbf{\Sigma}}^{-1}$	0.93	0.93	0.92	0.91	0.90	0.89
		(0.05)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
	$\hat{\mathbf{C}}^{aL}$	0.44	0.37	0.31	0.28	0.28	0.28
		(0.07)	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)
	$\hat{\mathbf{C}}^C$	0.00	0.00	0.01	0.10	0.20	0.19
		(0.00)	(0.00)	(0.03)	(0.12)	(0.04)	(0.03)
Conditional Heteroskedastic Errors							
$N = 10$	$\hat{\mathbf{\Sigma}}^{-1}$	0.92	0.93	0.95	0.96	0.96	0.96
		(0.05)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)
	$\hat{\mathbf{C}}^{aL}$	0.46	0.47	0.52	0.53	0.52	0.52
		(0.12)	(0.10)	(0.10)	(0.08)	(0.07)	(0.07)
	$\hat{\mathbf{C}}^C$	0.01	0.05	0.17	0.25	0.26	0.26
		(0.06)	(0.12)	(0.15)	(0.10)	(0.08)	(0.07)
$N = 20$	$\hat{\mathbf{\Sigma}}^{-1}$	0.92	0.92	0.91	0.90	0.89	0.90
		(0.04)	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
	$\hat{\mathbf{C}}^{aL}$	0.40	0.35	0.29	0.27	0.27	0.29
		(0.06)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)
	$\hat{\mathbf{C}}^C$	0.00	0.00	0.04	0.16	0.20	0.18
		(0.00)	(0.00)	(0.09)	(0.10)	(0.04)	(0.03)

[1] Monte Carlo standard errors are presented in parantheses below the Monte Carlo means of the criterion.

[2] $\hat{\mathbf{C}}^L$ denotes the LASSO, $\hat{\mathbf{C}}^{aL}$ the adaptive LASSO and $\hat{\mathbf{C}}^C$ the known DGP-class estimator when it is known that the data is generated by a VMA(1)-process.

[3] This table reports the error rates as fractions between committed Type II errors and the amount of true entries equal to zero on the lower triangle of the matrix.

Table 3.5: VAR(1) with Tridiagonal Precision Matrix – Norm Differences

		$T = 500$	$T = 1000$	$T = 2000$	$T = 3000$	$T = 4000$	$T = 5000$
Conditional Homoskedastic Errors							
$N = 10$	$\hat{\mathbf{\Sigma}}^{-1}$	1.27	0.91	0.68	0.59	0.53	0.49
		(0.17)	(0.12)	(0.08)	(0.07)	(0.06)	(0.06)
	$\hat{\mathbf{C}}^{L}$	1.16	0.87	0.66	0.56	0.51	0.47
		(0.17)	(0.11)	(0.09)	(0.07)	(0.07)	(0.06)
	$\hat{\mathbf{C}}^{aL}$	0.94	0.69	0.51	0.43	0.39	0.36
		(0.17)	(0.12)	(0.09)	(0.08)	(0.07)	(0.07)
	$\hat{\mathbf{C}}^{o}$	0.72	0.55	0.42	0.36	0.33	0.30
		(0.15)	(0.12)	(0.08)	(0.08)	(0.07)	(0.06)
	$\hat{\mathbf{C}}^{C}$	0.52	0.36	0.25	0.20	0.18	0.16
		(0.11)	(0.08)	(0.06)	(0.04)	(0.04)	(0.03)
$N = 20$	$\hat{\mathbf{\Sigma}}^{-1}$	3.07	2.30	1.84	1.63	1.50	1.40
		(0.24)	(0.17)	(0.12)	(0.11)	(0.10)	(0.10)
	$\hat{\mathbf{C}}^{L}$	2.59	2.08	1.75	1.55	1.42	1.32
		(0.23)	(0.19)	(0.12)	(0.11)	(0.10)	(0.10)
	$\hat{\mathbf{C}}^{aL}$	1.98	1.61	1.34	1.17	1.06	0.97
		(0.23)	(0.18)	(0.14)	(0.13)	(0.12)	(0.11)
	$\hat{\mathbf{C}}^{o}$	1.37	1.09	0.90	0.81	0.74	0.69
		(0.20)	(0.15)	(0.12)	(0.11)	(0.10)	(0.10)
	$\hat{\mathbf{C}}^{C}$	0.76	0.51	0.36	0.29	0.25	0.23
		(0.12)	(0.08)	(0.06)	(0.05)	(0.04)	(0.04)
Conditional Heteroskedastic Errors							
$N = 10$	$\hat{\mathbf{\Sigma}}^{-1}$	1.67	1.21	0.88	0.74	0.67	0.62
		(0.28)	(0.20)	(0.15)	(0.14)	(0.13)	(0.13)
	$\hat{\mathbf{C}}^{L}$	1.56	1.15	0.85	0.71	0.65	0.59
		(0.27)	(0.21)	(0.15)	(0.14)	(0.13)	(0.14)
	$\hat{\mathbf{C}}^{aL}$	1.35	0.99	0.73	0.61	0.55	0.51
		(0.28)	(0.22)	(0.17)	(0.16)	(0.14)	(0.15)
	$\hat{\mathbf{C}}^{o}$	1.22	0.92	0.69	0.58	0.52	0.48
		(0.29)	(0.23)	(0.17)	(0.16)	(0.14)	(0.14)
	$\hat{\mathbf{C}}^{C}$	1.08	0.79	0.57	0.48	0.42	0.38
		(0.26)	(0.21)	(0.15)	(0.14)	(0.12)	(0.13)
$N = 20$	$\hat{\mathbf{\Sigma}}^{-1}$	3.54	2.60	2.03	1.78	1.62	1.52
		(0.34)	(0.22)	(0.17)	(0.15)	(0.14)	(0.14)
	$\hat{\mathbf{C}}^{L}$	3.15	2.42	1.92	1.69	1.53	1.44
		(0.37)	(0.22)	(0.17)	(0.16)	(0.15)	(0.15)
	$\hat{\mathbf{C}}^{aL}$	2.53	1.95	1.53	1.33	1.19	1.12
		(0.35)	(0.25)	(0.21)	(0.20)	(0.18)	(0.18)
	$\hat{\mathbf{C}}^{o}$	1.96	1.51	1.20	1.06	0.95	0.89
		(0.33)	(0.27)	(0.22)	(0.21)	(0.18)	(0.18)
	$\hat{\mathbf{C}}^{C}$	1.66	1.20	0.88	0.72	0.63	0.57
		(0.29)	(0.24)	(0.18)	(0.16)	(0.13)	(0.14)

[1] Monte Carlo standard errors are presented in parantheses below the Monte Carlo means of the criterion.

[2] $\hat{\mathbf{\Sigma}}^{-1}$ denotes the inverse of the long-run covariance estimator, $\hat{\mathbf{C}}^{L}$ the LASSO, $\hat{\mathbf{C}}^{aL}$ the adaptive LASSO, $\hat{\mathbf{C}}^{o}$ the oracle and $\hat{\mathbf{C}}^{C}$ the known DGP-class estimator

Table 3.6: VAR(1) with Tridiagonal Precision Matrix – Type 1 Error Rates

		$T = 500$	$T = 1000$	$T = 2000$	$T = 3000$	$T = 4000$	$T = 5000$
Conditional Homoskedastic Errors							
$N = 10$	$\hat{\mathbf{C}}^L$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^{aL}$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^C$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
$N = 20$	$\hat{\mathbf{C}}^L$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^{aL}$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^C$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Conditional Heteroskedastic Errors							
$N = 10$	$\hat{\mathbf{C}}^L$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^{aL}$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^C$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
$N = 20$	$\hat{\mathbf{C}}^L$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^{aL}$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^C$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

[1] Monte Carlo standard errors are presented in parantheses below the Monte Carlo means of the criterion.

[2] $\hat{\mathbf{C}}^L$ denotes the LASSO, $\hat{\mathbf{C}}^{aL}$ the adaptive LASSO and $\hat{\mathbf{C}}^C$ the known DGP-class estimator when it is known that the data is generated by a VAR(1)-process.

[3] This table reports the error rates as fractions between committed Type II errors and the amount of true entries unequal to zero on the lower triangle of the matrix.

Table 3.7: VAR(1) with Tridiagonal Precision Matrix – Type 2 Error Rates

		$T = 500$	$T = 1000$	$T = 2000$	$T = 3000$	$T = 4000$	$T = 5000$
Conditional Homoskedastic Errors							
$N = 10$	$\hat{\mathbf{C}}^L$	0.89	0.91	0.91	0.90	0.89	0.89
		(0.07)	(0.05)	(0.05)	(0.05)	(0.06)	(0.06)
	$\hat{\mathbf{C}}^{aL}$	0.33	0.32	0.24	0.18	0.15	0.13
		(0.15)	(0.13)	(0.11)	(0.09)	(0.09)	(0.08)
	$\hat{\mathbf{C}}^C$	0.04	0.02	0.01	0.00	0.00	0.00
		(0.05)	(0.04)	(0.03)	(0.01)	(0.01)	(0.00)
$N = 20$	$\hat{\mathbf{C}}^L$	0.86	0.89	0.94	0.94	0.93	0.93
		(0.04)	(0.05)	(0.02)	(0.02)	(0.02)	(0.02)
	$\hat{\mathbf{C}}^{aL}$	0.27	0.29	0.36	0.31	0.27	0.24
		(0.08)	(0.13)	(0.06)	(0.05)	(0.05)	(0.05)
	$\hat{\mathbf{C}}^C$	0.02	0.01	0.00	0	0.00	0.00
		(0.02)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)
Conditional Heteroskedastic Errors							
$N = 10$	$\hat{\mathbf{C}}^L$	0.90	0.91	0.90	0.89	0.89	0.88
		(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
	$\hat{\mathbf{C}}^{aL}$	0.37	0.31	0.23	0.17	0.15	0.12
		(0.14)	(0.13)	(0.11)	(0.09)	(0.09)	(0.08)
	$\hat{\mathbf{C}}^C$	0.05	0.03	0.01	0.00	0.00	0.00
		(0.06)	(0.04)	(0.03)	(0.02)	(0.01)	(0.01)
$N = 20$	$\hat{\mathbf{C}}^L$	0.90	0.92	0.93	0.93	0.93	0.92
		(0.05)	(0.04)	(0.02)	(0.02)	(0.02)	(0.02)
	$\hat{\mathbf{C}}^{aL}$	0.37	0.36	0.34	0.29	0.26	0.24
		(0.12)	(0.11)	(0.06)	(0.06)	(0.05)	(0.05)
	$\hat{\mathbf{C}}^C$	0.02	0.01	0.00	0.00	0.00	0.00
		(0.02)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)

[1] Monte Carlo standard errors are presented in parantheses below the Monte Carlo means of the criterion.

[2] $\hat{\mathbf{C}}^L$ denotes the LASSO, $\hat{\mathbf{C}}^{aL}$ the adaptive LASSO and $\hat{\mathbf{C}}^C$ the known DGP-class estimator when it is known that the data is generated by a VAR(1)-process.

[3] This table reports the error rates as fractions between committed Type II errors and the amount of true entries equal to zero on the lower triangle of the matrix.

Table 3.8: VMA(1) with Erdös-Rényi Precision Matrix – Norm Differences

		$T = 500$	$T = 1000$	$T = 2000$	$T = 3000$	$T = 4000$	$T = 5000$
Conditional Homoskedastic Errors							
$N = 10$	$\hat{\mathbf{\Sigma}}^{-1}$	1.32	1.13	0.99	0.92	0.86	0.83
		(0.16)	(0.13)	(0.10)	(0.09)	(0.08)	(0.08)
	$\hat{\mathbf{C}}^{L}$	1.33	1.14	0.99	0.92	0.86	0.83
		(0.16)	(0.13)	(0.10)	(0.09)	(0.08)	(0.08)
	$\hat{\mathbf{C}}^{aL}$	1.31	1.11	0.96	0.88	0.82	0.79
		(0.20)	(0.15)	(0.11)	(0.10)	(0.09)	(0.09)
	$\hat{\mathbf{C}}^{o}$	0.91	0.78	0.68	0.62	0.58	0.56
		(0.17)	(0.13)	(0.10)	(0.09)	(0.08)	(0.08)
	$\hat{\mathbf{C}}^{C}$	3.07	2.48	1.92	1.75	1.66	1.59
		(0.34)	(0.37)	(0.20)	(0.11)	(0.09)	(0.10)
$N = 20$	$\hat{\mathbf{\Sigma}}^{-1}$	2.33	1.96	1.69	1.56	1.46	1.40
		(0.15)	(0.12)	(0.10)	(0.09)	(0.09)	(0.08)
	$\hat{\mathbf{C}}^{L}$	2.36	1.96	1.69	1.56	1.47	1.41
		(0.18)	(0.14)	(0.10)	(0.09)	(0.09)	(0.08)
	$\hat{\mathbf{C}}^{aL}$	2.39	1.84	1.54	1.42	1.33	1.28
		(0.53)	(0.26)	(0.12)	(0.10)	(0.10)	(0.09)
	$\hat{\mathbf{C}}^{o}$	1.31	1.13	0.97	0.89	0.82	0.79
		(0.17)	(0.14)	(0.11)	(0.09)	(0.09)	(0.08)
	$\hat{\mathbf{C}}^{C}$	4.01	3.25	2.46	2.04	1.89	1.80
		(0.43)	(0.36)	(0.38)	(0.18)	(0.12)	(0.11)
Conditional Heteroskedastic Errors							
$N = 10$	$\hat{\mathbf{\Sigma}}^{-1}$	1.43	1.21	1.05	0.95	0.89	0.85
		(0.24)	(0.21)	(0.19)	(0.16)	(0.15)	(0.14)
	$\hat{\mathbf{C}}^{L}$	1.43	1.21	1.05	0.96	0.90	0.85
		(0.25)	(0.21)	(0.19)	(0.16)	(0.15)	(0.14)
	$\hat{\mathbf{C}}^{aL}$	1.39	1.18	1.01	0.92	0.86	0.81
		(0.30)	(0.23)	(0.20)	(0.17)	(0.16)	(0.15)
	$\hat{\mathbf{C}}^{o}$	1.01	0.87	0.74	0.67	0.62	0.59
		(0.26)	(0.23)	(0.19)	(0.17)	(0.16)	(0.15)
	$\hat{\mathbf{C}}^{C}$	2.93	2.41	1.94	1.76	1.65	1.58
		(0.47)	(0.45)	(0.29)	(0.19)	(0.18)	(0.17)
$N = 20$	$\hat{\mathbf{\Sigma}}^{-1}$	2.49	2.04	1.74	1.59	1.49	1.42
		(0.20)	(0.16)	(0.15)	(0.13)	(0.12)	(0.11)
	$\hat{\mathbf{C}}^{L}$	2.46	2.03	1.73	1.59	1.49	1.42
		(0.21)	(0.17)	(0.15)	(0.13)	(0.12)	(0.11)
	$\hat{\mathbf{C}}^{aL}$	2.35	1.87	1.58	1.44	1.36	1.30
		(0.48)	(0.26)	(0.18)	(0.15)	(0.13)	(0.12)
	$\hat{\mathbf{C}}^{o}$	1.43	1.22	1.04	0.93	0.86	0.82
		(0.25)	(0.21)	(0.19)	(0.16)	(0.14)	(0.13)
	$\hat{\mathbf{C}}^{C}$	3.81	3.12	2.38	2.04	1.89	1.80
		(0.58)	(0.47)	(0.41)	(0.31)	(0.19)	(0.17)

[1] Monte Carlo standard errors are presented in parantheses below the Monte Carlo means of the criterion.

[2] $\hat{\mathbf{\Sigma}}^{-1}$ denotes the inverse of the long-run covariance estimator, $\hat{\mathbf{C}}^{L}$ the LASSO, $\hat{\mathbf{C}}^{aL}$ the adaptive LASSO, $\hat{\mathbf{C}}^{o}$ the oracle and $\hat{\mathbf{C}}^{C}$ the known DGP-class estimator

Table 3.9: VMA(1) with Erdös-Rényi Precision Matrix Precision Matrix – Type 1 Error Rates

		$T = 500$	$T = 1000$	$T = 2000$	$T = 3000$	$T = 4000$	$T = 5000$
Conditional Homoskedastic Errors							
$N = 10$	$\hat{\mathbf{C}}^L$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^{aL}$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^C$	0.13	0.04	0.00	0.00	0.00	0.00
		(0.09)	(0.03)	(0.01)	(0.00)	(0.00)	(0.00)
$N = 20$	$\hat{\mathbf{C}}^L$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^{aL}$	0.03	0.00	0.00	0.00	0.00	0.00
		(0.04)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^C$	0.09	0.05	0.01	0.00	0.00	0.00
		(0.03)	(0.03)	(0.02)	(0.00)	(0.00)	(0.00)
Conditional Heteroskedastic Errors							
$N = 10$	$\hat{\mathbf{C}}^L$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^{aL}$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^C$	0.11	0.03	0.00	0.00	0.00	0.00
		(0.09)	(0.04)	(0.02)	(0.00)	(0.00)	(0.00)
$N = 20$	$\hat{\mathbf{C}}^L$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^{aL}$	0.02	0.00	0.00	0.00	0.00	0.00
		(0.03)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^C$	0.08	0.04	0.01	0.00	0.00	0.00
		(0.04)	(0.03)	(0.02)	(0.00)	(0.00)	(0.00)

[1] Monte Carlo standard errors are presented in parantheses below the Monte Carlo means of the criterion.

[2] $\hat{\mathbf{C}}^L$ denotes the LASSO, $\hat{\mathbf{C}}^{aL}$ the adaptive LASSO and $\hat{\mathbf{C}}^C$ the known DGP-class estimator when it is known that the data is generated by a VMA(1)-process.

[3] This table reports the error rates as fractions between committed Type I errors and the amount of true entries unequal to zero on the lower triangle of the matrix.

Table 3.10: VMA(1) with Erdös-Rényi Precision Matrix – Type 2 Error Rates

		$T = 500$	$T = 1000$	$T = 2000$	$T = 3000$	$T = 4000$	$T = 5000$
Conditional Homoskedastic Errors							
$N = 10$	$\hat{\mathbf{C}}^L$	0.96	0.97	0.98	0.98	0.98	0.98
		(0.04)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
	$\hat{\mathbf{C}}^{aL}$	0.52	0.50	0.52	0.48	0.46	0.43
		(0.15)	(0.16)	(0.12)	(0.11)	(0.11)	(0.11)
	$\hat{\mathbf{C}}^C$	0.04	0.16	0.29	0.27	0.26	0.27
		(0.11)	(0.20)	(0.12)	(0.09)	(0.09)	(0.11)
$N = 20$	$\hat{\mathbf{C}}^L$	0.86	0.95	0.96	0.96	0.96	0.97
		(0.17)	(0.08)	(0.01)	(0.02)	(0.02)	(0.02)
	$\hat{\mathbf{C}}^{aL}$	0.37	0.43	0.38	0.35	0.35	0.41
		(0.24)	(0.10)	(0.05)	(0.07)	(0.10)	(0.12)
	$\hat{\mathbf{C}}^C$	0.03	0.05	0.20	0.27	0.23	0.21
		(0.05)	(0.08)	(0.16)	(0.06)	(0.04)	(0.04)
Conditional Heteroskedastic Errors							
$N = 10$	$\hat{\mathbf{C}}^L$	0.97	0.97	0.98	0.98	0.98	0.98
		(0.04)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
	$\hat{\mathbf{C}}^{aL}$	0.53	0.51	0.51	0.47	0.45	0.43
		(0.15)	(0.15)	(0.12)	(0.11)	(0.11)	(0.11)
	$\hat{\mathbf{C}}^C$	0.08	0.18	0.28	0.27	0.27	0.28
		(0.16)	(0.19)	(0.13)	(0.10)	(0.11)	(0.11)
$N = 20$	$\hat{\mathbf{C}}^L$	0.90	0.95	0.96	0.96	0.96	0.97
		(0.14)	(0.06)	(0.02)	(0.02)	(0.02)	(0.02)
	$\hat{\mathbf{C}}^{aL}$	0.42	0.43	0.37	0.35	0.37	0.42
		(0.19)	(0.09)	(0.06)	(0.09)	(0.12)	(0.12)
	$\hat{\mathbf{C}}^C$	0.06	0.10	0.23	0.26	0.23	0.21
		(0.12)	(0.15)	(0.14)	(0.06)	(0.04)	(0.04)

[1] Monte Carlo standard errors are presented in parantheses below the Monte Carlo means of the criterion.

[2] $\hat{\mathbf{C}}^L$ denotes the LASSO, $\hat{\mathbf{C}}^{aL}$ the adaptive LASSO and $\hat{\mathbf{C}}^C$ the known DGP-class estimator when it is known that the data is generated by a VMA(1)-process.

[3] This table reports the error rates as fractions between committed Type II errors and the amount of true entries equal to zero on the lower triangle of the matrix.

Table 3.11: VAR(1) with Erdös-Rényi Precision Matrix Structure – Norm Differences

		$T = 500$	$T = 1000$	$T = 2000$	$T = 3000$	$T = 4000$	$T = 5000$
Conditional Homoskedastic Errors							
$N = 10$	$\hat{\mathbf{\Sigma}}^{-1}$	0.90	0.66	0.51	0.45	0.42	0.40
		(0.12)	(0.08)	(0.07)	(0.06)	(0.06)	(0.05)
	$\hat{\mathbf{C}}^{L}$	0.88	0.66	0.51	0.45	0.42	0.40
		(0.11)	(0.09)	(0.07)	(0.06)	(0.06)	(0.06)
	$\hat{\mathbf{C}}^{aL}$	0.79	0.60	0.46	0.41	0.38	0.36
		(0.12)	(0.10)	(0.08)	(0.07)	(0.06)	(0.06)
	$\hat{\mathbf{C}}^{o}$	0.59	0.47	0.38	0.35	0.33	0.31
		(0.12)	(0.10)	(0.08)	(0.07)	(0.07)	(0.06)
	$\hat{\mathbf{C}}^{C}$	0.52	0.36	0.26	0.21	0.18	0.16
		(0.10)	(0.07)	(0.05)	(0.04)	(0.04)	(0.04)
$N = 20$	$\hat{\mathbf{\Sigma}}^{-1}$	1.93	1.30	0.95	0.81	0.73	0.68
		(0.17)	(0.10)	(0.07)	(0.05)	(0.05)	(0.04)
	$\hat{\mathbf{C}}^{L}$	1.69	1.16	0.87	0.77	0.71	0.67
		(0.14)	(0.08)	(0.07)	(0.06)	(0.05)	(0.04)
	$\hat{\mathbf{C}}^{aL}$	1.17	0.83	0.65	0.59	0.54	0.5
		(0.13)	(0.09)	(0.07)	(0.06)	(0.05)	(0.05)
	$\hat{\mathbf{C}}^{o}$	0.86	0.67	0.54	0.48	0.45	0.43
		(0.13)	(0.09)	(0.07)	(0.06)	(0.05)	(0.05)
	$\hat{\mathbf{C}}^{C}$	0.88	0.62	0.43	0.35	0.30	0.27
		(0.11)	(0.07)	(0.05)	(0.04)	(0.03)	(0.03)
Conditional Heteroskedastic Errors							
$N = 10$	$\hat{\mathbf{\Sigma}}^{-1}$	1.14	0.83	0.63	0.54	0.49	0.46
		(0.19)	(0.14)	(0.13)	(0.11)	(0.10)	(0.10)
	$\hat{\mathbf{C}}^{L}$	1.12	0.82	0.62	0.54	0.49	0.46
		(0.19)	(0.14)	(0.13)	(0.11)	(0.10)	(0.11)
	$\hat{\mathbf{C}}^{aL}$	1.00	0.74	0.56	0.48	0.44	0.42
		(0.20)	(0.17)	(0.15)	(0.12)	(0.12)	(0.12)
	$\hat{\mathbf{C}}^{o}$	0.82	0.63	0.49	0.43	0.40	0.37
		(0.20)	(0.17)	(0.15)	(0.12)	(0.12)	(0.12)
	$\hat{\mathbf{C}}^{C}$	0.85	0.62	0.45	0.37	0.33	0.30
		(0.21)	(0.16)	(0.13)	(0.10)	(0.10)	(0.10)
$N = 20$	$\hat{\mathbf{\Sigma}}^{-1}$	2.33	1.57	1.12	0.95	0.85	0.79
		(0.28)	(0.17)	(0.12)	(0.11)	(0.09)	(0.09)
	$\hat{\mathbf{C}}^{L}$	2.06	1.42	1.05	0.92	0.82	0.77
		(0.27)	(0.18)	(0.13)	(0.11)	(0.09)	(0.09)
	$\hat{\mathbf{C}}^{aL}$	1.52	1.08	0.83	0.72	0.65	0.61
		(0.26)	(0.20)	(0.15)	(0.12)	(0.11)	(0.11)
	$\hat{\mathbf{C}}^{o}$	1.17	0.90	0.69	0.60	0.55	0.52
		(0.22)	(0.18)	(0.14)	(0.12)	(0.10)	(0.10)
	$\hat{\mathbf{C}}^{C}$	1.36	0.98	0.71	0.58	0.51	0.46
		(0.23)	(0.19)	(0.14)	(0.12)	(0.10)	(0.10)

[1] Monte Carlo standard errors are presented in parantheses below the Monte Carlo means of the criterion.

[2] $\hat{\mathbf{\Sigma}}^{-1}$ denotes the inverse of the long-run covariance estimator, $\hat{\mathbf{C}}^{L}$ the LASSO, $\hat{\mathbf{C}}^{aL}$ the adaptive LASSO, $\hat{\mathbf{C}}^{o}$ the oracle and $\hat{\mathbf{C}}^{C}$ the known DGP-class estimator

Table 3.12: VAR(1) with Erdös-Rényi Precision Matrix – Type 1 Error Rates

		$T = 500$	$T = 1000$	$T = 2000$	$T = 3000$	$T = 4000$	$T = 5000$
Conditional Homoskedastic Errors							
$N = 10$	$\hat{\mathbf{C}}^L$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^{aL}$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^C$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
$N = 20$	$\hat{\mathbf{C}}^L$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^{aL}$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^C$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Conditional Heteroskedastic Errors							
$N = 10$	$\hat{\mathbf{C}}^L$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^{aL}$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^C$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
$N = 20$	$\hat{\mathbf{C}}^L$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^{aL}$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\hat{\mathbf{C}}^C$	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

[1] Monte Carlo standard errors are presented in parantheses below the Monte Carlo means of the criterion.

[2] $\hat{\mathbf{C}}^L$ denotes the LASSO, $\hat{\mathbf{C}}^{aL}$ the adaptive LASSO and $\hat{\mathbf{C}}^C$ the known DGP-class estimator when it is known that the data is generated by a VAR(1)-process.

[3] This table reports the error rates as fractions between committed Type I errors and the amount of true entries unequal to zero on the lower triangle of the matrix.

Table 3.13: VAR(1) with Erdös-Rényi Precision Matrix – Type 2 Error Rates

		$T = 500$	$T = 1000$	$T = 2000$	$T = 3000$	$T = 4000$	$T = 5000$
Conditional Homoskedastic Errors							
$N = 10$	$\hat{\mathbf{C}}^L$	0.95	0.97	0.97	0.97	0.97	0.96
		(0.04)	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)
	$\hat{\mathbf{C}}^{aL}$	0.45	0.44	0.38	0.31	0.26	0.23
		(0.15)	(0.15)	(0.10)	(0.10)	(0.09)	(0.09)
	$\hat{\mathbf{C}}^C$	0.13	0.11	0.11	0.10	0.10	0.09
		(0.06)	(0.05)	(0.04)	(0.03)	(0.03)	(0.03)
$N = 20$	$\hat{\mathbf{C}}^L$	0.94	0.93	0.92	0.95	0.97	0.97
		(0.02)	(0.02)	(0.03)	(0.04)	(0.02)	(0.01)
	$\hat{\mathbf{C}}^{aL}$	0.41	0.31	0.25	0.30	0.35	0.32
		(0.05)	(0.05)	(0.09)	(0.12)	(0.06)	(0.04)
	$\hat{\mathbf{C}}^C$	0.22	0.22	0.23	0.24	0.24	0.25
		(0.05)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
Conditional Heteroskedastic Errors							
$N = 10$	$\hat{\mathbf{C}}^L$	0.96	0.97	0.97	0.97	0.96	0.96
		(0.04)	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)
	$\hat{\mathbf{C}}^{aL}$	0.48	0.44	0.37	0.31	0.26	0.23
		(0.15)	(0.14)	(0.11)	(0.1)	(0.09)	(0.09)
	$\hat{\mathbf{C}}^C$	0.14	0.12	0.11	0.10	0.10	0.10
		(0.07)	(0.05)	(0.04)	(0.03)	(0.03)	(0.03)
$N = 20$	$\hat{\mathbf{C}}^L$	0.94	0.94	0.94	0.96	0.97	0.97
		(0.02)	(0.03)	(0.03)	(0.03)	(0.02)	(0.01)
	$\hat{\mathbf{C}}^{aL}$	0.41	0.33	0.31	0.34	0.34	0.32
		(0.08)	(0.09)	(0.12)	(0.10)	(0.06)	(0.05)
	$\hat{\mathbf{C}}^C$	0.22	0.23	0.24	0.24	0.25	0.25
		(0.05)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)

[1] Monte Carlo standard errors are presented in parantheses below the Monte Carlo means of the criterion.

[2] $\hat{\mathbf{C}}^L$ denotes the LASSO, $\hat{\mathbf{C}}^{aL}$ the adaptive LASSO and $\hat{\mathbf{C}}^C$ the known DGP-class estimator when it is known that the data is generated by a VAR(1)-process.

[3] This table reports the error rates as fractions between committed Type II errors and the amount of true entries equal to zero on the lower triangle of the matrix.

CHAPTER 4

Robustness of Financial Volatility Networks to the Exclusion of Systemic Nodes

This chapter is based on the identically entitled working paper

In recent years, network analysis has become and increasingly popular tool to analyse large panels of time series. In particular, there is an evergrowing literature which utilizes network analysis to gauge systemicness of firms and sectors in financial markets, especially during the Financial Crisis of 2007–2010. A common feature in this literature is that Lehman Brothers is excluded from the sampled data due to its bankruptcy in mid-September 2008. However, it is well known that omitting central nodes in the analysis of networks induces bias in network measures. Using this as a starting point, I empirically assess how the exclusion of Lehman Brothers' stock from the sample affects estimation outcomes for volatility networks by estimating the widely applied long-run variance decomposition network of Diebold and Yilmaz (2014) based on a commonly used panel of 101 major U.S. firms' stock price volatilities where I explicitly in- and exclude Lehman Brothers. This allows me to gauge the effects Lehman Brothers' stock has on the commonly used From- and To-degree network measures. I find that the To-degree is heavily affected by the exclusion of Lehman Brothers whereas the From-degree seems to be only minorly affected. These results hold on a firm-specific and aggregated sector level for a sparse and non-sparse VAR-representation of the data.

4.1 Introduction

Arbitrage pricing theory and capital asset pricing models suggest that volatility, like returns, are governed by a systematic and an idiosyncratic part. The systematic part is assumed to be directly quantifiable by common, observed factors (such as the market return in the classic capital asset pricing model), whereas the idiosyncratic part is unobserved. Moreover, classic financial theory suggests that the systematic part is non-diversifiable. That is, it has to be taken as given and one can not protect oneself against it. In contrast, the idiosyncratic part is commonly assumed to be perfectly diversifiable as more and more assets are considered, cf. Chamberlain and Rothschild (1983). In particular, it is argued that different assets offset their idiosyncratic risks. However, recent studies suggest that the idiosyncratic part is not diversifiable (see, inter alia, Gabaix, 2011; Acemoglu et al., 2012) and, therefore, even if an increasing number of assets is considered, the idiosyncratic part is non-negligible. Thus, if the idiosyncratic part is assumed to be perfectly diversifiable, the true risk propagation mechanism is neglected. In addition, the actual risk in the financial system is erroneously assessed, which can have severe economic consequences. For example, market regulations would not target the actual risk and, as a consequence, are not able to mitigate the effects of a crisis.

To better gauge the risk transmission channels in large financial systems, network[1] analysis has become a widely applied tool to measure the inherent idiosyncratic risk, cf. the pioneering work of Diebold and Yilmaz (2009, 2011, 2012, 2014). These authors analyze the idiosyncratic risk of return and volatility series by constructing a network based on the Wold-representation of the underlying data. In particular, the network is constructed based on the variance decomposition scheme of Pesaran and Shin (1998). Inspired by this work, Billio et al. (2012) investigate idiosyncratic risk in the finance and insurance sector in monthly return series of hedge funds, banks, brokers and insurance companies. Moreover, Demirer et al. (2015) analyze idiosyncratic risk in global bank networks and Bostanci and Yilmaz (2015) analyze idiosyncratic risk in global sovereign credit risk networks. More recently, Barigozzi and Brownlees (2017) proposed a methodology for quantifying the idiosyncratic risk in high dimensional financial series based on the Generalized Dynamic Factor Model (GDFM) of Forni et al. (2015, 2017).

However, all of this work considers sub-samples of the true underlying financial networks. In particular, usually 100 U.S. stocks with the majority taken from the S&P 100 are sampled. Moreover, while assessing systemic risk during the Financial Crisis of 2007–2010, these studies drop Lehman Brothers from their sample, presumably because Lehman Brothers filed for bankruptcy on September, 15^{th} 2008 and the stock stopped being traded two days later. But doing so implies a major problem to the analysis of financial networks and networks in general. In particular, Kolaczyk (2017, Chapter 3) provides simulation results and theoretical considerations

[1] A network consists of nodes (e.g. firms) and edges (connections amongst the nodes). Thus, edges can be seen as transmission channels and nodes with many edges can be regarded as important roles in transmitting the idiosyncratic risk

which show that omitting nodes while sampling a network induces a bias on the estimates of basic network measures such as the overall and the node specific degrees. Moreover, it is also shown that the implied biases depend on a multitude of factors, for example the topology of the network, the sampling scheme and the chosen measure itself. Based on this, this study looks into the effects of Lehman Brothers' omission from the sample has on results currently put forward in the empirical literature on financial networks.

Following Diebold and Yilmaz (2014), this work is based on volatility series rather than return series. This is due to the fact that volatilities are sensitive to crisis periods and that volatilities are often regarded as a tracking device for investor fear. Thus, volatilities are a suitable candidate to track and quantify risk transmission channels. I consider a panel of daily stock prices of 100 U.S. firms plus those of Lehman Brothers, which is observed from 01.01.2007 up to and including 17.09.2008, and I compute for these stock prices their latent volatility series with the range measure of Parkinson (1980). That is, the volatility of a firm's stock is calculated as a scaled difference between the highest and lowest log-price of the asset on any given day in the sample. Following Barigozzi and Brownlees (2017), the systematic part is then, similar to the classic capital asset pricing model, approximated by a linear function of the market and SPDR sector volatilities.[2,3] Thus, the idiosyncratic risk is represented by the residual series of these regressions and is the main object of interest for the analysis. In order to quantify the idiosyncratic risk and its transmission channels, the Long Run Variance Decomposition Network (LVDN) of Diebold and Yilmaz (2014) is used. The LVDN is based on the forecast error variation of variable i due to shocks to variable j. Thus, the network is completely defined by a VMA(∞) representation of the panel of residual series.

I find that omitting Lehman Brothers from the sample does not alter the qualitative properties of the network. In particular, the financial sector is still the most influential sector in the network. However, quantitative results change to some extend. In particular, the results suggest that the From- and To-degrees associated with firms in the financial sector tend to be underestimated whereas the effects of non-financial firms tend to be overestimated. The first finding is most likely due to the fact an important financial institution is deleted from the sample and the latter due to the fact that non-financial firms pick up some of the connections originally originating from Lehman Brothers.

The remainder of this chapter is structured as follows. Section 4.2 introduces basic terminology used in the analysis of networks and introduces the LVDN based on the VMA(∞)-representation of a second-order stationary stochastic process. Section 4.3 outlines a simple approach to model volatility LVDNs and Section 4.4 describes the estimation approach. Section 4.5 discusses the empirical findings and Section 4.6 concludes.

[2]Market and sector volatilities are also computed by the high-low range measure of Parkinson (1980). Moreover, each firm is assigned to one of the nine SPDR sectors (the Real Estate sector is excluded due to data availability).

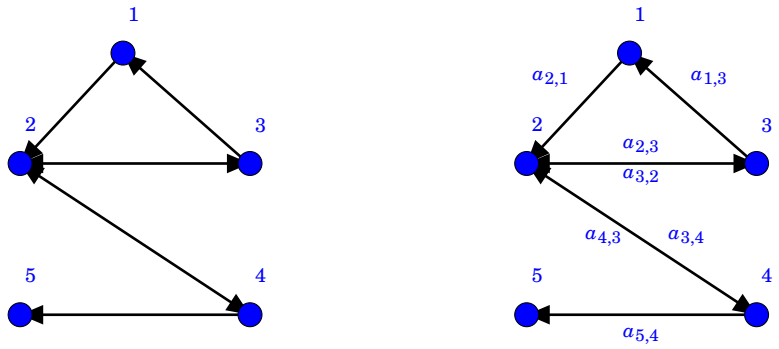
[3]Note that Barigozzi and Brownlees (2017) use the GDFM to first estimate volatilities and subsequently apply the GDFM again to extract the observed systematic factors. For simplicity, I stick to the usual measure of volatility described above.

4.2 The Long-Run Variance Decomposition Network

Networks based on Time Series Data For the purpose of this chapter it suffices to consider directed networks. A directed network or graph $\mathcal{G}$ is defined as the tuple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, 2, ..., N\}$, $N \in \mathbb{N}$, denotes the set of nodes or vertices (individuals, firms or stock returns to give some examples) and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the edge set (connections between nodes). The edge set $\mathcal{E}$ can also be represented by use of an adjacency matrix $\mathbf{A}$. In case of directed networks $\mathbf{A}$ is not symmetric, the diagonal elements are equal to one and the off-diagonal entries having non-zero value $a_{i,j}$ if and only if there is an edge from node j and i. If it holds that $a_{i,j} = 1$ for all $i \neq j$ then $\mathcal{G}$ is called an unweighted directed network. In contrast, the network is called weighted directed if at least two distinct non-zero entries $a_{i,j}$ and $a_{i',j'}$ of the adjacency matrix satisfy $a_{i,j} \neq a_{i',j'}$ and $a_{i,j}$ denotes the weight of the edge from node j and i. Note that $a_{i,j} = 1$ is still a possibility and that it can be that $a_{i,j} \neq 0$ and $a_{j,i} = 0$ or $a_{i,j} \neq 0$ and $a_{j,i} \neq 0$ with $a_{i,j} \neq a_{j,i}$. Figure 4.1 illustrates such networks.

Figure 4.1: Examples of Directed Networks

The left figure shows an unweighted directed and the right figure a weighted directed network as indicated by the weights $a_{i,j}$ next to each edge. In case where connections go from node i to node j and vice versa, both weights are depicted. Note that in case of an unweighted network no weights are displayed in the figure since they do not provide further information. Nodes are labelled with numerals and lines correspond to edges.



For a given network $\mathcal{G}$, one might be interested in identifying important nodes or the average connectedness, i.e. average amount of (weighted) edges per node in the network. A basic, yet important measure to answer such questions is the degree of a node i which measures the amount of edges attached to it. Obviously, for an undirected network there are two such measures: a) the From-degree (or In-degree) summing all (weighted) edges ending in node i and b) the To-degree (or Out-degree) summing all (weighted) edges leaving node i. That is, the From-degree measures how strongly node i is directly influenced by all other nodes in the network and the To-degree measures how strongly other nodes in the network are directly influenced by node i. Finally, the Total-degree measures the average amount of (weighted) edges per node in the network by

averaging either over all From- or all To-degrees[4]. A network is said to be more connected than another if its Total-degree is larger. Note that these concepts are measures of systemic risk as in Diebold and Yilmaz (2014) and Barigozzi and Hallin (2017) amongst others, and their definition will be formalized in the next section after the particular network - LVDN - is introduced in the section below.

4.2.1 Construction of the LVDN

In order to measure interconnectedness within economic systems such as financial markets, Diebold and Yilmaz (2014) propose the usage of the LVDN which is based on the second-order stationary VMA(∞)-representation of a random process $\mathbf{y}_t$,

$$\mathbf{y}_t = \sum_{l=0}^{\infty} \mathbf{\Theta}_l \boldsymbol{\eta}_{t-l}, \quad \boldsymbol{\eta}_t \sim w.n.(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{\eta}}), \tag{4.1}$$

which summarizes all aspects of connectedness. In particular, $\mathbf{\Sigma}_{\boldsymbol{\eta}}$ and $\mathbf{\Theta}_0$ contain all contemporaneous effects, and $\{\mathbf{\Theta}_1, \mathbf{\Theta}_2, ...\}$ all dynamic aspects of connectedness.

Based on (4.1), Diebold and Yilmaz (2014) propose to compute the adjacency matrix of the LVDN based on the generalized variance decomposition of Pesaran and Shin (1998),

$$d_{i,j}^H = \frac{\tilde{\sigma}_{\boldsymbol{\eta},jj} \sum_{h=0}^{H-1} (\mathbf{e}_i^\top \mathbf{\Theta}_h \mathbf{\Sigma}_{\boldsymbol{\eta}} \mathbf{e}_j)^2}{\sum_{h=0}^{H-1} (\mathbf{e}_i^\top \mathbf{\Theta}_h \mathbf{\Sigma}_{\boldsymbol{\eta}} \mathbf{\Theta}_h^\top \mathbf{e}_i)}, \tag{4.2}$$

where $\mathbf{e}_i$ is a $N \times 1$-vector with 1 on the i–th entry, and zeros everywhere else, H is the forecast horizon and $\tilde{\sigma}_{\boldsymbol{\eta},jj}$ denotes the (j,j)-th entry of $\mathbf{\Sigma}_{\boldsymbol{\eta}}^{-1}$. Since the innovations in (4.1) are not necessarily orthogonal, sums of forecast error variance contributions do not necessarily sum to one, making direct interpretation of the entries difficult. To circumvent this, the adjacency matrix of the LVDN, $\mathbf{W}^H$, has (i,j)-th entry

$$w_{i,j}^H = 100 \frac{d_{i,j}^H}{\sum_{j=1}^N d_{i,j}^H} \tag{4.3}$$

where the scaling by 100 is used to express percentages. Note that in practice the LVDN depends on the forecast horizon H and there is no reason why the LVDN should be the same for different H. Equation (4.3) also implies that the LVDN is a *weighted, directed* network.

Since I use the LVDN to identify important nodes and sectors, appropriate measures need to be defined in order identify such nodes. For that, I will make use of the commonly applied To- and the From-degree. Formally, these two measures are defined as

$$\delta_j^{\text{To}} = \sum_{\substack{i=1 \\ i \neq j}}^{N} w_{i,j}^H, \; j = 1, ..., N \quad \text{and} \quad \delta_i^{\text{From}} = \sum_{\substack{j=1 \\ j \neq i}}^{N} w_{i,j}^H, \; i = 1, ..., N. \tag{4.4}$$

[4]Of course, both averages are equal since any edge leaving one node (and, therefore, contributing to the To-degree of this node) must end at another node (and, therefore, contributing to the From-degree of this second node).

Note that a node with a high To-degree influences other nodes to a high degree, whereas a node with a high From-degree is influenced by other nodes to a high degree. A node for which both, the in- and out-degree are large, is greatly influenced by other nodes and also greatly influences other nodes as well. This naturally leads to the following definition of systemic nodes.

Definition 4.1. *A node i in the LVDN, defined by the adjacency matrix $\mathbf{W}^H$ with (i,j)-th entry given by* (4.3)*, is said to be systemic if its From- or To-degrees are relatively large compared to the same degrees of the other nodes.*

Based on this definition, I will call a node systemic when either of its two degree measures is large relative to the degree measures of the other nodes. Moreover, one note about this definition is of order. Of course, the term *large* has different meaning to each individual researcher. However, under this definition it still prevails that the nodes with the highest From- and To-degrees are central to the system in spreading and attracting shocks quicker than other nodes since there are more/stronger connections associated with that node. Also worth mentioning at this point is that Brownlees and Mesters (2017) recently proposed a different measure for systemicness of a node based on whether the node is granular or not. However, this approach is not pursued in this study since most current work on volatility networks uses the notion of systemicness in Definition 4.1.

4.3 A Factor Approach for Volatility

Since this chapter focuses on identifying systemic nodes in the financial market and how such results might change under inclusion/exclusion of nodes, the systematic risk exposure needs to be extracted before the systemic risk can be assessed. Following Barigozzi and Brownlees (2017), I approximate the systematic part by the market, respectively sector volatilities. Based on this, I model the log-volatility, denoted by $\ln\sigma^2_{i,t}$, of stock i on day t as

$$\ln\sigma^2_{i,t} = \alpha_i + \beta_i \ln\sigma^2_{m,t} + \gamma_i \ln\sigma^2_{s_i,t} + \epsilon_{i,t}, \quad i \in \{1,...,N\}, t \in \{1,...,T\} \tag{4.5}$$

where $\sigma^2_{m,t}$ and $\sigma^2_{s_i,t}$ denote the volatility of the S&P500 index, respectively the volatility of the SPDR sectoral index of the S&P500[5] to which firm i belongs. Consequently, the systemic risk, associated with the financial market, is captured by the error-terms $\epsilon_{i,t}$, which will be the subject of the forthcoming analysis.

Computing the (partial) autocorrelation function of the residual series $\{\hat{\epsilon}_{i,t} : i = 1,...,N, t = 1,...,T\}$ indicates that there is still autocorrelation present. Therefore, I approximate the residual series by a stationary VAR(1)-process[6]. which directly links to the VMA(∞)-representation needed

[5]The SPDR sectors are: Materials (XLB), Utilities (XLU), Energy (XLE), Industrials (XLI), Technology (XLK), Consumer Staples (XLP), Health Care (XLV), Financials (XLF) and Consumer Discretionary (XLY).

[6]As will become apparent in Section 5, my samples consists of 431 observations. Thus, fitting higher order VAR-processes is unreliable in such small samples.

for constructing the LVDN. Denote $\boldsymbol{\epsilon}_t = (\epsilon_{1,t}, ..., \epsilon_{N,t})^\top$ and let $\mathbf{B}$ be a $N \times N$-matrix satisfying $\det(\mathbf{I}_N - \mathbf{B}z) \neq 0$ for any $z \in \mathbb{C}$ such that $|z| \leq 1$. The residual series is then approximated by

$$\boldsymbol{\epsilon}_t = \mathbf{B}\boldsymbol{\epsilon}_{t-1} + \mathbf{u}_t, \quad t = 1, ...T \tag{4.6}$$

where $\mathbf{u}_t$ is assumed to be a white-noise error term with full rank covariance matrix $\boldsymbol{\Sigma}_\mathbf{u}$, which is not necessarily diagonal in general. After netting out the market and sector volatilities, most interdependencies among the entries in $\hat{\boldsymbol{\epsilon}}_t$ are expected to vanish. Therefore, I assume that $\max_j \sum_{i=1}^N \mathbb{1}\{(\mathbf{B})_{i,j} \neq 0\} = o(N)$, $\max_i \sum_{j=1}^N \mathbb{1}\{(\mathbf{B})_{i,j} \neq 0\} = o(N)$ and $\max_j \sum_{i=1}^N \mathbb{1}\{(\boldsymbol{\Sigma}_\mathbf{u}^{-1})_{i,j} \neq 0\} = o(N)$. Hence, estimating (4.6) by some regularization technique, like the adaptive LASSO, is appropriate to enforce this (unknown) sparsity structure.

Since (4.6) is assumed to be a second order stationary VAR(1)-model, it possesses a VMA(∞)-representation as in (4.1) and, consequently, the adjacency matrix of a LVDN can be computed as described in (4.2) by inverting the VAR(1)-processes. Hence, based on this representation the subsequent analysis will be carried out.

4.4 Estimation

4.4.1 Measuring Volatility via Realized Range and Extraction of the Residual Series

Since the aim of this chapter is to estimate volatility networks and volatility is an unobserved quantity, it needs to be estimated. To do so, I follow Diebold and Yilmaz (2015) and Barigozzi and Brownlees (2017) and utilize the high-low range volatility measure of Parkinson (1980) to quantify firm i's stock price volatility on day t,

$$\hat{\sigma}_{i,t}^2 = \frac{(\ln H_{i,t} - \ln L_{i,t})^2}{4\ln 2}, \tag{4.7}$$

where $H_{i,t}$, respectively $L_{i,t}$ denotes the highest, respectively lowest price of firm i's stock on day t[7]. Note that Parkinson (1980) derives this volatility measure under the assumption that the stock-price of firm i follows a geometric Brownian motion.

Even though the above high-low volatility measure is simple in its nature and more advanced measures have been proposed in the literature (see, inter alia, Andersen et al., 2003; Barndorff-Nielsen et al., 2008, 2011), several studies show that it performs well in terms of small bias and variance (see, inter alia, Bali and Weinbaum, 2005; Martens and van Dijk, 2007; Brownlees and Gallo, 2010).

Moreover, note that this volatility measure does not distinguish between the diffusion and the jump part of volatility. This is due to the following reasons. First, the scope of the present study is to assess how the exclusion of central nodes affects currently available results. To be

[7] Of course, this approach is also used to compute the market and sector volatilities.

able to compare the results to the literature, I follow Diebold and Yilmaz (2015) and Barigozzi and Brownlees (2017) and use the above. Second, if one is interested in central nodes and sectors only, then disentangling them is not necessary since only the overall volatility of each firm is of interest. If, however, one wants to see which part of the volatility matters most, then one is well advised trying to disentangle them.

Finally, (4.5) is estimated by least squares and the residual series $\hat{\boldsymbol{\epsilon}}_t$ is obtained in the usual way.

4.4.2 Estimation of the LVDN

In line with the literature, I estimate the VAR(1)-process in (4.6) by OLS and use the sample covariance, $T^{-1}\sum_t \hat{\mathbf{u}}_t\hat{\mathbf{u}}_t^\top$, which is based on the least squares residuals of (4.6) as an estimator for the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{u}}$ of the errors $\mathbf{u}_t$ in (4.6). However, given the dimensionality and the number of observations in the used sample these estimates can be unreliable. To circumvent this issue I will also consider a sparse representation of the VAR(1)-process in (4.6). In particular, I estimate Equation (4.6) by minimizing the penalized quadratic loss

$$\min_{\mathbf{A}} \sum_{t=1}^{T} \|\hat{\boldsymbol{\epsilon}}_t - \mathbf{B}\hat{\boldsymbol{\epsilon}}_{t-1}\|_F^2 + \lambda_{\mathbf{B}} \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{|b_{ij}|}{|\tilde{b}_{ij}|}, \quad t = 1,...T \tag{4.8}$$

where $\lambda_{\mathbf{B}} \geq 0$ is the penalization parameter and $\tilde{b}_{ij}$ a pre-estimator (an initial LASSO estimator in this case) for the (i,j)-th entry of $\mathbf{B}$. The above minimization problem is solved via the adaptive LASSO of Zou (2006). I choose the adaptive LASSO over the classic LASSO of Tibshirani (1996) because it is able to perform consistent model selection, cf. Zou (2006). Furthermore, I choose the penalization parameter by minimizing the BIC over a grid of possible values. Note that I also carry out the analysis when $\mathbf{B}$ is assumed to not be sparse for comparison reasons.

Lastly, an estimator for $\boldsymbol{\Sigma}_{\mathbf{u}}$ is needed. Given the VAR-parameters, obtained in the previous step, the residuals $\hat{\mathbf{u}}_t$ can be computed as $\hat{\mathbf{u}}_t = \hat{\boldsymbol{\epsilon}}_t - \hat{\boldsymbol{\epsilon}}_{t-1}\hat{\mathbf{B}}$ for $t = 1,..,T$ and their inverse covariance matrix is inferred via the `space`-algorithm of Peng et al. (2009). This algorithm estimates N single regressions by means of LASSO from which the partial correlations, $\rho_{i,j}$, between $\hat{u}_i$ and $\hat{u}_j$ are computed and afterwards their inverse covariance matrix. This is possible due to the fact that the (i,j)-th entry of the inverse covariance matrix is proportional, in a known fashion, to the partial correlation between $\hat{u}_i$ and $\hat{u}_j$:

$$\hat{\rho}_{i,j} = -\frac{\tilde{\sigma}_{\hat{\mathbf{u}},ij}}{\sqrt{\tilde{\sigma}_{\hat{\mathbf{u}},ii}\tilde{\sigma}_{\hat{\mathbf{u}},jj}}}, \tag{4.9}$$

see Lauritzen (1996). Here, $\tilde{\sigma}_{\hat{\mathbf{u}},ij}$ denotes the (i,j)-th entry of $\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{u}}}^{-1}$. Note that the `space`-algorithm provides estimates for $\rho_{i,j}$ and $\tilde{\sigma}_{\hat{\mathbf{u}},ij}$, $i,j = 1,...,N$.

4.5 Empirical Analysis

In this section, empirical results will be presented. Before doing so, I will briefly discuss the data set used for this empirical study.

4.5.1 Data Description

The data in this empirical study consists of 101 U.S. stocks taken from the S&P500 index (a list of the included series and their tickers can be found in Table 4.5). The majority of these stocks are listed in the S&P100.[8] The data is sampled on a daily frequency from January, 1^{st} 2007 up to and including September, 17^{th}, 2008, resulting in 431 observations over time for each series and was downloaded from Yahoo-Finance and from the Wharton Research Data Base in case of Lehman Brothers. Note that September, 17^{th}, 2008 is the last day on which prices for Lehman Brothers are observed after the company filed for bankruptcy.

Two remarks about the used data set are of order at this point. First, I explicitly include Lehman Brothers. The inclusion of Lehman Brothers is crucial since all current studies on volatility networks exclude it, thereby missing a probably important node in their data set. As mentioned in the Introduction, exclusion of nodes induces sampling error in the network estimates and, consequently, in the network To- and From-degree measures. Thus, this allows me to gauge to what extend the currently prevailing results in the empirical volatility network might be contaminated by omission of key nodes. Second, and closely connected to this issue is the fact that I, as the majority of the current literature (see, inter alia, Barigozzi and Brownlees, 2017; Brownlees and Mesters, 2017; Barigozzi and Brownlees, 2017, and references therein), consider such a data set. One can argue that only using these 101 stocks one still misses 399 stocks to obtain the full S&P500. Therefore, one still has the problem of biased network measures. However, the S&P100's market capitalization makes up about 50% of the market capitalization in the U.S. equity market. Therefore, it seems plausible to argue that the remaining firms, which are left out from the data set, play a minor role and that their bias effect on the estimated network does not alter the results by too much.

4.5.2 Results

In this Section, results will be presented. First, I am going to discuss how the omission of Lehman Brothers' stock affects individual network measures and, afterwards, how aggregate measures such as From- and To-degrees of firms aggregated in sectors are affected. Note that throughout this section I set the forecast horizon to $H = 10$.[9]

[8] Note that due to the way the S&P100 is constructed. In fact firms can enter and leave the S&P100 on a regular basis since a panel of experts decides which firms are listed. Therefore, not all of the firms in the considered data set were permanent constituents of the S&P100. However, 90 of the considered firms were permanent constituents of the S&P100 during the sampling period.

[9] $H = 10$ is chosen since the current Basel accord requires a 10-day value at risk assessment. Moreover, given the construction of the LVDN, it can be expected that exclusion of Lehman Brothers from the sample does not alter the

Effect of Lehman Brothers on Firm-Specific Network Measures The results for the sample including Lehman Brothers are presented in Table 4.1. Table 4.2 displays results for the same sample exluding Lehman Brothers. First, we can note that in both cases when Lehman

Table 4.1: 10 Largest Network Measures for Firms, including Lehman Brothers

To-Degree			From-Degree		
Non-Sparse VAR					
Ticker	Sector	Degree	Ticker	Sector	Degree
SPG	XLF	58.84	GS	XLF	9.72
LEH	XLF	54.36	GE	XLI	9.69
AIG	XLF	45.33	MS	XLF	9.40
C	XLF	42.69	MSFT	XLK	9.35
MS	XLF	26.30	ABT	XLV	9.32
NEM	XLB	17.03	BK	XLF	9.20
DVN	XLE	13.18	NOV	XLE	9.19
COF	XLF	11.84	HPQ	XLK	9.18
GS	XLF	11.10	COP	XLE	9.06
PG	XLP	11.06	MDT	XLV	9.00
Sparse VAR					
Ticker	Sector	Degree	Ticker	Sector	Degree
LEH	XLF	36.87	GS	XLF	9.58
SPG	XLF	30.28	MS	XLF	9.33
AIG	XLF	28.68	GE	XLI	9.19
C	XLF	28.06	ABT	XLV	8.65
MS	XLF	14.70	HPQ	XLK	8.60
GE	XLI	11.09	MSFT	XLK	8.51
DVN	XLE	10.24	NEM	XLB	8.47
LMT	XLI	8.99	AIG	XLF	8.45
COF	XLF	8.73	NOV	XLE	8.29
GS	XLF	8.57	AEP	XLU	8.25

[1] Sector abbreviations are: Materials (XLB), Energy (XLE), Financials (XLF), Industrials (XLI), Technology (XLK), Consumer Staples (XLP), Utilities (XLU), Health Care (XLV) and Consumer Discretionaries (XLY).
[2] The firm tickers can be found in Table 4.5 in Appendix 4.A.

Table 4.2: 10 Largest Network Measures for Firms, excluding Lehman Brothers

To-Degree			From-Degree		
Non-Sparse VAR					
Ticker	Sector	Degree	Ticker	Sector	Degree
SPG	XLF	54.44	GS	XLF	9.72
AIG	XLF	45.21	GE	XLI	9.59
C	XLF	40.00	MSFT	XLK	9.34
MS	XLF	28.75	MS	XLF	9.32
NEM	XLB	18.25	ABT	XLV	9.22
DVN	XLE	14.55	NOV	XLE	9.20
COF	XLF	14.11	BK	XLF	9.15
GE	XLI	13.78	HPQ	XLK	9.15
PG	XLP	12.03	COP	XLE	9.06
ABT	XLV	11.42	MDT	XLV	8.96
Sparse VAR					
Ticker	Sector	Degree	Ticker	Sector	Degree
AIG	XLF	36.99	GE	XLI	9.25
MS	XLF	23.74	GS	XLF	9.20
SPG	XLF	22.27	MS	XLF	8.94
C	XLF	16.70	MSFT	XLK	8.74
COF	XLF	10.34	NEM	XLB	8.60
USB	XLF	10.12	HPQ	XLK	8.52
GS	XLF	9.94	ABT	XLV	8.46
ABT	XLV	9.94	NOV	XLE	8.20
GE	XLI	9.69	AEP	XLU	8.11
LMT	XLI	9.29	AIG	XLF	8.10

[1] Sector abbreviations are: Materials (XLB), Energy (XLE), Financials (XLF), Industrials (XLI), Technology (XLK), Consumer Staples (XLP), Utilities (XLU), Health Care (XLV) and Consumer Discretionaries (XLY)
[2] The firm tickers can be found in Table 4.5 in Appendix 4.A.

Brothers is omitted or not the financial sector plays a dominant role in propagating shocks since in either case the top 10 From-degrees are filled with financial firms. In terms of attracting shocks, the picture is somewhat different since the top 10 firms come from a variety of sectors with no clear pattern. However, the financial sector still has the most firms in the top 10 To-degrees. Note that the top 25 firms with the highest From-degrees all have a From-degree measure of 8.5 or higher and include several financial firms. Moreover, Lehman Brothers plays a major role in distributing shock through the network by being ranked in the top 2 most "To-connected" firms.

A closer look at the top 10 firms of the To-degree in the sample from which Lehman Brothers is excluded reveals that the majority of firms stays in the top 10 for both the sparse and the

effects for other forecast horizons H.

non-sparse VAR setting. However, their ranking shuffles. Moreover, General Electric (GE) and Abbott Laboratories enter the top 10 in both VAR settings once Lehman Brothers is excluded. In case of the non-sparse VAR setting Goldman Sachs (GS) leaves the top 10 and for the sparse VAR-setting, Devon Newport (DVN) leaves and USB enters the top 10. Thus, there is some movement in the top 10 due to the exclusion of Lehman Brothers from the sample.

Focusing now on the actual To-degrees themselves one quickly notices that they change by a large margin after the deletion of Lehman Brothers from the sample. All financial firms which stay within the top 10 have larger changes in their To-degrees than non-financial firms. Moreover, the changes are by far larger in case of the sparse VAR-setting. For example, Citigroup's To-degree decreases by about 11 percentage points in case of a sparse VAR but only by about 2.7 percentage points in case of a non-sparse VAR-setting. These findings suggest that the exclusion of Lehman Brothers from the data has substantial effects on the estimated To-degrees, irrespective of the used VAR-setting and the changes are most striking for financial firms. This is most likely due to the fact that these firms pick up effects which are originally attributed to Lehman Brothers. Finally, there is no clear direction in which the changes occurr since both increases and decreases can be observed.

When the From-degrees are considered, the picture changes. In particular, the composition of the top 10 firms remains the same, only the ordering of the firms changes. The actual From-degrees themselves change by less than 0.1 percentage point in case of a non-sparse VAR. For the sparse VAR setting, however, there some larger changes but still not as pronounced as for the To-degrees. In particular, the financial firms' From-degrees change at most by 0.5 percentage points and the non-financial firms' degrees by less than 0.1 percentage points with the exception of American Electric Power (AEP). Thus, it seems that the exclusion of Lehman Brothers does not affect the From-degrees as much as the To-degrees. But given the fact that the From-degree of Lehman Brothers is relatively small (they are not in the top 10) there is also not much which can carry over after exclusion.

Effect of Lehman Brothers on Sector Measures In this Section I consider the nine SPDR sectors and their associated degree measures. Similar to the From- and To-degrees for individual nodes we can also compute From- and To-degrees on a sector level. That is, we collect all nodes which belong to the same SPDR sectors and take the average of their From- and To-degrees, respectively. The results for this exercise can be found in Tables 4.3 for the data set which includes Lehman Brothers and in Table 4.4 for the data set excluding Lehman Brothers.

Starting with the To-degree measure of the nine sectors, one can see that the financial sector clearly dominates the other sectors by having a To-degree two to three times as large as the second ranked sector. Moreover, we can observe that the the To-degree decreases by about 2 percentage points in case of a non-sparse VAR specification and by about 1 percentage point for the sparse VAR specification. For all other sectors the To-degree changes by at most 1 percentage point. In particular, the materials and the health care sectors have both similar changes. All

other sectors' To-degrees change by even less than 0.5 percentage points. However, even these small changes shuffle the sector orderings after Lehman Brothers got deleted. Concluding,

Table 4.3: SPDR Sectors ranked by degree measures, Lehman Brothers included

To-Degree		From-Degree	
Non-Sparse VAR			
Sector	Degree	Sector	Degree
XLF	20.28	XLF	8.52
XLI	6.53	XLV	8.21
XLB	6.47	XLU	8.10
XLV	6.33	XLE	8.06
XLE	5.88	XLB	8.01
XLU	5.76	XLK	7.89
XLK	5.33	XLI	7.80
XLP	4.93	XLP	7.48
XLY	4.76	XLY	6.97
Sparse VAR			
Sector	Degree	Sector	Degree
XLF	13.77	XLF	7.49
XLI	5.70	XLU	6.80
XLU	5.30	XLE	6.56
XLB	5.05	XLV	6.24
XLK	4.88	XLB	6.21
XLE	4.88	XLI	6.17
XLY	4.54	XLK	6.03
XLV	4.33	XLY	5.05
XLP	4.13	XLP	4.84

[1] Sector abbreviations are: Materials (XLB), Energy (XLE), Financials (XLF), Industrials (XLI), Technology (XLK), Consumer Staples (XLP), Utilities (XLU), Health Care (XLV) and Consumer Discretionaries (XLY)

Table 4.4: SPDR Sectors ranked by degree measures, Lehman Brothers excluded

To-Degree		From-Degree	
Non-Sparse VAR			
Sector	Degree	Sector	Degree
XLF	18.05	XLF	8.45
XLB	7.49	XLV	8.14
XLV	7.17	XLU	8.10
XLI	6.92	XLE	8.02
XLE	6.11	XLB	7.95
XLU	5.97	XLK	7.89
XLK	5.72	XLI	7.80
XLP	5.27	XLP	7.43
XLY	5.12	XLY	6.93
Sparse VAR			
Sector	Degree	Sector	Degree
XLF	12.56	XLF	7.53
XLI	6.18	XLU	6.75
XLB	5.81	XLE	6.47
XLU	5.55	XLV	6.45
XLV	5.33	XLB	6.36
XLY	4.82	XLK	6.23
XLK	4.79	XLI	6.23
XLP	4.75	XLY	5.15
XLE	4.69	XLP	4.88

[1] Sector abbreviations are: Materials (XLB), Energy (XLE), Financials (XLF), Industrials (XLI), Technology (XLK), Consumer Staples (XLP), Utilities (XLU), Health Care (XLV) and Consumer Discretionaries (XLY)

these findings suggest that results on financial networks reported in the current literature are trustworthy in a qualitative sense. That is, in agreement with intuition the financial sector plays the key role in propagating and attracting shocks in the U.S. financial system. However, quantitative statements about their effects might be in err given my findings. In particular, the above results suggest that reported findings might be biased downwards for the financial sector and upwards for the remaining sectors.

4.6 Conclusions

In this chapter I investigated to what extend the omission of Lehman Brothers from a sample of 101 U.S. firms' stock price volatilities affects results in the widely applied LVDN based on a VAR(1) representation of the data. I find that the central role of the financial sector and its constituents remains unaltered when Lehman Brothers is deleted from the sample. However, I also find that the estimated To-degree measure can change to a large extend. For example, Citigroup's To-degree decreases by about eleven percentage points after Lehman Brothers is excluded from the sample. Moreover, the results suggest that the From-degree measures are less affected by the exclusion of Lehman Brothers. Nevertheless, they do indeed change as well. Thus, the exclusion of important firms, such as Lehman Brothers, from the analysis can yield misleading findings in empirical studies, both on a firm-specific and on an aggregated sector-level.

Appendix 4.A Tables

Table 4.5: Data Description

Ticker	Company Name	Ticker	Company Name
XLB		XLU	
DD	Du Pont	AEP	American Electric Power
DOW	Dow Chemicals	EXC	Exelon
FCX	Freeport-McMoran	IDA	IdaCorp.
MON	Monsanto	MGEE	MGE Energy
MLM	Martin Marietta Materials	PNW	Pinnacle West Capital Corp.
NEM	Newmont Mining Corp.	SO	Southern Company
SEE	Sealed Air Corp.	WR	Westar Energy
VMC	Vulcan Materials Company	XEL	Xcel Energy
XLI		XLK	
BA	Boeing Company	AAPL	Apple
CAT	Caterpillar	ACN	Accenture plc
EMR	Emerson Electric	CSCO	Cisco Systems
FDX	FedEx	EBAY	eBay
GD	General Dynamics	EMC	EMC
GE	General Electric	HPQ	Hewlett-Packard
HON	Honeywell Intl.	IBM	IBM
LMT	Lockheed Martin	INTC	Intel
MMM	3M Company	MSFT	Microsoft
NSC	Norfolk Southern	ORCL	Oracle
RTN	Raytheon	QCOM	QUALCOMM
UNP	Union Pacific	TXN	Texas Instruments
UPS	United Parcel Service	T	AT&T

Continued on next page

Table 4.5 – continued from previous page

Ticker	Company Name	Ticker	Company Name
UTX	United technologies	VZ	Verizon
	XLE		XLP
APA	Apache	CL	Colgate-Palmolive
APC	Anadarko Petroleum	COST	Costco
COP	ConocoPhillips	CVS	CVS Caremark
CVX	Chevron	KO	The Coca Cola Company
DVN	Devon Energy	MDLZ	Mondelez International
HAL	Halliburton	MO	Altria
NOV	National Oilwell Varco	Pep	PepsiCo
OXY	Occidental Petroleum	PG	Procter & Gamble
SLB	Schlumberger Ltd.	WMT	Wal-Mart Stores
XOM	Exxon Mobile	–	–
	XLY		XLV
AMZN	Amazon.com	ABT	Abbott Laboratories
CMCSA	Comcast	AMGN	Amgen
DIS	Walt Disney	BAX	Baxter International
F	Ford Motor	BMY	Bristol-Myers Squibb
FOXA	Twenty-First Century Fox	GILD	Gilead Sciences
HD	Home Depot	JNJ	Johnson & Johnson
LOW	Lowes	LLY	Lilly (Eli) & Co.
MCD	McDonalds	MDT	Medtronic
NKE	Nike	MRK	Merck & Co.
SBUX	Starbucks	PFE	Pfizer
TGT	Target	UNH	United Health
TWX	Time Warner	–	–
	XLF		
AIG	AIG	JPM	JPMorgan Chase
ALL	Allstate	LEH	Lehman Brothers
AXP	American Express Co.	MET	MetLife
BAC	Bank of America	MS	Morgan Stanley
BK	Bank of New York	SPG	Simon Property
C	Citigroup	USB	U.S. Bankcorp.
COF	Capital One Financial	WFC	Wells Fargo
GS	Goldman Sachs		

[1] The sector abbreviations are: Materials (XLB), Energy (XLE), Financials (XLF), Industrials (XLI), Technology (XLK), Consumer Staples (XLP), Utilities (XLU), Health Care (XLV) and Consumer Discretionary (XLY)

BIBLIOGRAPHY

ACEMOGLU, D., V. M. CARVALHO, A. OZDAGLAR, AND A. TAHBAZ-SALEHI (2012): "The Network Origins of Aggregate Fluctuations," *Econometrica*, 80, 1977–2016.

ANDERSEN, T. G., T. BOLLERSLEV, F. X. DIEBOLD, AND P. LABYS (2003): "Modelling and Forecasting Realized Volatility," *Econometrica*, 71, 579–625.

ANDREWS, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

——— (1993): "Tests for Parameter Instability and Structural Change With Unknown Change Point," *Econometrica*, 61, 821–856.

ANDREWS, D. W. K. AND J. C. MONAHAN (1992): "An Improved heteroskedasticty and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica*, 60, 953–966.

ANTOINE, B. AND O. BOLDEA (2015): "Inference in Linear Models with Structural Changes and Mixed Identification Strength," SFU Working Paper 15-5.

——— (2017): "Efficient Inference with Time-Varying Information and the New Keynesian Phillips Curve," SFU Working Paper, http://www.sfu.ca/ baa7/research/AntoineBoldea201708.pdf.

BAI, J. AND P. PERRON (1998): "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica*, 66, 47–78.

BALI, T. G. AND D. WEINBAUM (2005): "A Comparative Study of Alternative Extreme-Value Volatility Estimators," *The Journal of Futures Markets*, 25, 873–892.

BANERJEE, O., L. EL GHAOUI, AND A. D'ASPREMONT (2008): "Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data," *The Journal of Machine Learning Research*, 9, 485–516.

BANERJEE, S. AND S. GHOSAL (2013): "Bayesian Estimation of a Sparse Precision Matrix," *arXiv:1309.1754v2*.

BARIGOZZI, M. AND C. BROWNLEES (2017): "NETS: Network Estimation For Time Series," Working Paper.

BARIGOZZI, M. AND M. HALLIN (2017): "A Network Analysis of the Volatility of High Dimensional Financial Series," *Journal of the Royal Statistical Society – Series C*, 581–605.

BARNDORFF-NIELSEN, O. E., P. R. HANSEN, A. LUNDE, AND N. SHEPHARD (2008): "Designing Realized Kernels to Measure the ex post Variation of Equity Prices in the Presence of Noise," *Econometrica*, 76, 1481–1536.

——— (2011): "Multivariate Realised Kernels: Consistent Positive Semi-Definite Estimators of the Covariation of Equity Prices with Noise and Non-Synchronous Trading," *Journal of Econometrics*, 162, 149–169.

BICKEL, P. J. AND E. LEVINA (2008): "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, 36, 199–227.

BILLIO, M., M. GETMANSKY, A. W. LO, AND L. PELIZZON (2012): "Econometric measures of connectedness and system risk in the finance and insurance sector," *Journal of Financial Economics*, 104, 535–559.

BOLDEA, O., A. CORNEA, AND A. HALL (2017): "Bootstrapping Structural Change Tests," Working Paper.

BOSTANCI, G. AND K. YILMAZ (2015): "How connected is the Global Sovereign Credit Risk Network?" Koc University-TUSIAD Economic Research Forum Working Paper No:1515.

BOYD, S. AND L. VANDENBERGHE (2004): *Convex Optimization*, Cambridge University Press, UK.

BREGMAN, L. M. (1967): "The Relaxation Method for Finding the Common Point of Convex Sets and its Application to the Solution of Problems in Convex Programming," *USSR Computational Mathematics and Mathematical Physics*, 7, 191–204.

BREIMAN, L. (1996): "Heuristics of Instability and Stabilization in Model Selection," *Annals of Statistics*, 24, 2350–2383.

BROWNLEES, C. AND G. M. GALLO (2010): "Comparison of Volatility Measures: A Risk Management Perspective," *Journal of financial Econometrics*, 8, 29–56.

BROWNLEES, C. AND G. MESTERS (2017): "Detecting Granular Time Series in Large Panels," Working Paper.

CAI, T. T., H. LI, W. LIU, AND J. XIE (2014): "Joint Estimation of Multiple High-dimensional Precision Matrices," Tech. rep., working Paper.

CAI, T. T., W. LIU, AND X. LUO (2011): "A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation," *Journal of the American Statistical Association*, 106, 594–607.

CAI, T. T., W. LIU, AND H. H. ZHOU (2012): "Estimating Sparse Precision Matrix: Optimal Rates of Convergence and Adaptive Estimation," *arXiv:1212.2882v1*.

CAI, T. T., Z. REN, AND H. H. ZHOU (2016): "Estimating Structured High-Dimensional Covaraince and Precision Matrices: Optimal Rates and Adaptive Estimation," *Electronic Journal of Statistics*, 10, 1–59.

CANER, M. AND B. HANSEN (2001): "Threshold Autoregression with a Unit Root," *Econometrica*, 69, 1555–1596.

CANER, M. AND B. E. HANSEN (2004): "Instrumental Variable Estimation of a Threshold Model," *Econometric Theory*, 20, 813–843.

CHAMBERLAIN, G. AND M. ROTHSCHILD (1983): "Arbitrage, Factor Structure, and Mean-Variance Analysis on large Asset Markets," *Econometrica*, 51, 1281–1304.

CHAN, K. (1993): "Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Autoregressive Model," *Annals of Statistics*, 21, 520–533.

CHEN, X., M. XU, AND W. B. WU (2013): "Covariance and Precision Matrix Estimation for High-Dimenional Time Series," *The Annals of Statistics*, 41, 2994–3021.

CHRISTIANO, L., M. EICHENBAUM, AND S. REBELO (2011): "When is the Government Spending Multiplier Large?" *Journal of Political Economy*, 119, 78–121.

DAVIDSON, R. AND J. G. MACKINNON (1998): "Graphical Methods for Investigating the Size and Power of Hypothesis Tests," *The Manchester School*, 66, 1–26.

DAVIES, R. B. (1977): "Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternative," *Biometrika*, 64, 247–254.

DEMIRER, M., F. X. DIEBOLD, L. LIU, AND K. YILMAZ (2015): "Estimating Global Bank Network Connectedness," Koc University-TUSIAD Economic Research Forum Working Paper No:1512.

DEMPSTER, A. (1972): "Covariance Selection," *Biometrics*, 28, 157–175.

DIEBOLD, F. X. AND K. YILMAZ (2009): "Measuring Financial Asset Return and Volatility Spillovers, With Application to Global Equity Markets," *Economic Journal*, 119, 158–171.

——— (2011): *Equity Market Spillovers in the Americas*, Santiago: Bank of Chile Central Banking Series, 199–214.

——— (2012): "Better to Give than to Receive: Forecast-Based Measurement of Volatility Spillovers," *International Journal of Forecasting*, 28, 57–66.

——— (2014): "On the Network Topology of Variance Decompositions: Measuring the Connectedness of Financial Firms," *Journal of Econometrics*, 182, 119–134.

——— (2015): *Measuring the Dynamics of Global Business Cycle Connectedness*, Oxford University Press, Chapter 5.

EGGERTSSON, G. (2010): "What Fiscal Policy Is Effective at Zero Interest Rates?" in *NBER Macroeconomic Annual*, ed. by D. Acemogly and M. Woodford, University of Chicago Press.

EL KAROUI, N. (2008): "Operator Norm Consistent Estimation of Large-Dimensional Sparse Covariance Matrices," *The Annals of Statistics*, 36, 2717–2756.

FAN, J., Y. FENG, AND Y. WU (2009): "Network Exploration via the Adaptive LASSO and SCAD Penalties," *The Annals of Applied Statistics*, 3, 521–541.

FAN, J., Y. LIAO, AND H. YUAN (2016): "An overview of the estimation of large covariance and precision matrices," *The Econometrics Journal*, 19, C1–C32.

FORNI, M., M. HALLIN, M. LIPPI, AND P. ZAFFARONI (2015): "Dynamic Factor Models with Inifinite-Dimensional Factor Spaces: One Sided Representations," *Journal of Econometrics*, 185, 359–371.

——— (2017): "Dynamic Factor Models with Inifinite-Dimensional Factor Spaces: Asymptotic Analysis," *Journal of Econometrics*, 199, 74–92.

FRIEDMAN, J., T. HASTIE, AND R. TIBSHIRANI (2008): "Sparse Inverse Covaraince Estimation with the Graphical LASSO," *Biostat*, 9, 432–441.

GABAIX, X. (2011): "The Granular Origins of Aggregate Fluctuations," *Econometrica*, 79, 733–772.

GONZALO, J. AND J.-Y. PITARAKIS (2002): "Estimation and Model Selection Based Inference in Single and Multiple Threshold Models," *Journal of Econometrics*, 110, 319–352.

——— (2006): "Threshold Effects in Cointegrating Regressions," *Oxford Bulletin of Economics and Statistics*, 68, 813–833.

GONZALO, J. AND M. WOLF (2005): "Subsampling Inference in Threshold Autoregressive Models," *Journal of Econometrics*, 127, 201–224.

HALL, A. R., S. HAN, AND O. BOLDEA (2012): "Inference Regarding Multiple Structural Changes in Linear Models with Endogenous Regressors," *Journal of Econometrics*, 170, 281–302.

HANSEN, B. (2016): "Regression Kink with an Unknown Threshold," *Journal of Business and Economic Statistics*, forthcoming.

HANSEN, B. E. (1996): "Inference when a Nuisance Parameter is Not Identified under the Null Hypothesis," *Econometrica*, 64, 413–430.

——— (1999): "Threshold Effects in Non-Dynamic Panels: Estimation, Testing, and Inference," *Journal of Econometrics*, 93, 345–368.

——— (2000): "Sample Splitting and Threshold Estimation," *Econometrica*, 68, 575–603.

——— (2011): "Threshold Autoregression in Economics," *Statistics and Its Interface*, 4, 123–127.

HORN, R. A. AND C. R. JOHNSON (2013): *Matrix Analysis*, Cambridge University Press, 2nd ed.

KOLACZYK, E. (2017): *Topics at the Frontier of Statistics and Network Analysis: (Re)Visiting the Foundations*, Cambridge: Cambridge University Press.

KOURTELLOS, A., T. STENGOS, AND C. TAN (2015): "Structural Threshold Regression," *Econometric Theory*, 1–34.

LAM, C. AND J. FAN (2009): "Sparsistency and Rates of Convergence in Large Covariance Matrix Estimation," *The Annals of Statistics*, 37, 4254–4278.

LAURITZEN, S. L. (1996): *Graphical Models (Oxford Statistical Science Series)*, Oxford University Press, USA.

LEDOIT, O. AND M. WOLF (2003): "Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection," *Journal of Empirical Finance*, 10, 603–621.

LEE, S., H. PARK, M. SEO, AND Y. SHIN (2014): "A Contribution to the Reinhart and Rogoff Debate: not 90% but maybe 30%," *CEMMAP Working Paper CWP 39-14*.

LEE, S., M. H. SEO, AND Y. SHIN (2011): "Testing for Threshold Effects in Regression Models," *Journal of the American Statistical Association*, 106, 220–231.

LI, H. AND J. GUI (2006): "Gradient Directed Regularization for Sparse Gaussian Concentration Graphs with AApplication to Inference of Genetic Networks," *Biostatistics*, 7, 302–317.

LI, J., P. STOICA, AND Z. WANG (2003): "On Robust Capon Beamforming and Diagonal Loading," *IEEE Transactions on Signal Processing*, 51, 1702–1715.

MAGNUS, J. R. AND H. NEUDECKER (1979): "The Commutation Matrix: Some Properties and Applications," *Annals of Statistics*, 7, 381–394.

MAGNUSSON, L. AND S. MAVROEIDIS (2014): "Identification Using Stability Restrictions," *Econometrica*, 82, 1799–1851.

MAMMEN, E. (1993): "Bootstrap and Wild Bootstrap for High-Dimensional Linear Models," *Annals of Statistics*, 21, 255–285.

MARTENS, M. AND D. VAN DIJK (2007): "Measuring Volatility with the Realized Range," *Journal of Econometrics*, 138, 181–207.

MEINSHAUSEN, N. AND P. BÜHLMANN (2006): "High dimensional graphs and variable selection with the lasso," *Annals of Statistics*, 34, 1436–1462.

NEWEY, W. K. AND K. D. WEST (1994): "Automatic Lag Selection in Covariance Matrix Estimation," *Review of Economic Studies*, 61, 631–654.

PARKINSON, M. (1980): "The Extreme Value Method for estimating the Variance of the Rate of Return," *The Journal of Business*, 53, 61–65.

PENG, J., P. WANG, N. ZHOU, AND J. ZHU (2009): "Partial Correlation Estimation by Joint Sparse Regression Models," *Journal of the American Statistical Association*, 104, 735–746.

PESARAN, M. H. AND Y. SHIN (1998): "Generalized Impulse Response Analysis in Linear Multivariate Models," *Economics Letters*, 58, 17–29.

PHILLIPS, P. C., Y. SUN, AND S. JIN (2007): "Long Run Variance Estimation and Robust Regression Testing using Sharp Origin Kernels with no Truncation," *Journal of Statistical Planning and Inference*, 137, 985–1023.

RAVIKUMAR, P., M. J. WAINWRIGHT, G. RASKUTTI, AND B. YU (2011): "High-Dimensional Covariance Estimation by minimizing ℓ_1-penalized log-determinant Divergence," *Electronic Journal of Statistics*, 5, 935–980.

REINHART, C. M. AND K. S. ROGOFF (2010): "Growth in a Time of Debt," *American Economic Review: Papers and Proceedings*, 100, 573–578.

SEGAL, E., N. FRIEDMAN, N. KAMINSKI, A. REGEV, AND D. KOLLER (2005): "From Signatures to Models: Understanding Cancer using Microarrays," *Nature Genetics*, 37, S38–S45.

SEO, M. AND O. LINTON (2007): "A Smoothed Least Squares Estimator for Threshold Regression Models," *Journal of Econometrics*, 141, 704–735.

TALIH, M. (2003): "Markov Random Fields on Time-Varying Graphs, with an Application to Portfolio Selection," Ph.D. thesis, Yale University, ProQuest LLC, Ann Arbor, MI.

TIBSHIRANI, R. (1996): "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society. Series B*, 58, 267–288.

TONG, H. (1990): *Nonlinear Time Series - A Dynamical System Approach*, Oxford: Clarendon Press.

YU, P. AND P. PHILLIPS (2014): "Threshold Regression with Endogeneity," Cowles Foundation Working Paper 1966.

YUAN, M. AND Y. LIN (2007): "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35.

ZOU, H. (2006): "The adaptive LASSO and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.