High Dimensional Time Series — New Techniques and Applications

Zur Erlangung des akademischen Grades
Doktor der Wirtschaftswissenschaften
(Dr. rer. pol)

bei der Fakultät für Wirtschaftswissenschaften
Karlsruher Institut für Technologie (KIT)

genehmigte
Dissertation

von
M.Sc. Chong Liang

Tag der mündlichen Prüfung:	18.07.2018
Referentin:	Prof. Dr. Melanie Schienle
Korreferent:	Prof. Dr. Enno Mammen

Karlsruhe 31.07.2018

Acknowledgement

I would like to express my deep gratitude to my supervisor Prof. Dr. Melanie Schienle for her tremendous help from 2011 when I came to Germany. I benefit a lot from her lectures and our cooperations. During these years, she keeps on helping improve this Dissertation from technical details to the texts. The harmonious research environment she creates for the whole team is a valuable treasure for me. I would also like to express my warm thanks to Prof. Dr. Enno Mammen for his help with this dissertation and the time traveling from Heidelberg for my doctoral defense.

During these years, many experts in econometrics and statistics from different universities have helped with the Dissertation. To name a few: Bernd Droge (Humboldt-Universität zu Berlin), Alexey Onatskiy (University of Cambridge), Weibiao Wu (University of Chicago), Qiwei Yao (LSE), Kyusang Yu (Konkuk University), Rongmao Zhang (Zhejiang University). I would also like to express thanks to my colleagues in these years, especially Ruihong Huang for improving my programming skills.

Finally I would like to thank my wife Shi Chen for her understanding and love during the past few years. I am so lucky to know her in Germany, who makes my life more than research. My parents Jianwen Liang and Meidong Li receive my deepest gratitude and love for their dedication, who raised me with a love of science and supported me in all my pursuits.

Contents

1 Introduction

The past decade witnessed the rapid development of high dimensional statistics in deterministic design. High dimensional time series analysis, due to the time dependency, still faces several theoretical challenges. Among the time series models, the Vector Error Correction Model (VECM) is especially complicated because of the non-stationary components. The classical estimation strategies (e.g. Johansen's approach) fail to provide consistent estimates for dimensions larger than three. Moreover, it is impossible to apply existing statistical methods to determine VECM in high dimensions. This dissertation aims at providing feasible regularized methods, which can determine and estimate high dimensional VECM with robust statistical properties. The detailed analysis is divided into three parts contained in Chapters 2-4. I develop new tailored Lasso-type methods and prove their statistical properties. From the application side these techniques are highly valuable for appropriately treating complex potentially non-stationary systems not only in economics, finance but also in weather and climate systems. I also illustrate this for portfolios of Credit Default Swaps in a banking-sovereign network (Chapter 3). In Chapter 5, I provide a detailed empirical study on a high-frequency portfolio where new high-dimensional techniques allow to account for liquidity effects through the Limit Order Book in a very detailed way. With this the new spillover channels in the system can be identified.

In Chapter 2, I introduce the idea treating cointegration rank selection in VECM as a Lasso-type variable selection problem. Such an approach relies on the QR-decomposition of the least squared estimator, which provides the pre-estimators for both the possible orthonormal basis spanning the cointegrating space and the corresponding loading matrix. Therefore, the inner products between the non-stationary components and the different basis become the regressors in the Lasso step. In the true model, only r (the rank) out of m (the dimension) basis have non-zero loading coefficients. To detect the r important basis, I construct weights from the pre-estimator of the loading matrix and apply adaptive Lasso for the purpose of consistent model selection. The number of non-zero columns in the penalized estimator for the loading matrix is exactly the cointegration rank. Lag selection is relatively easier since it only includes stationary components. Therefore, the adaptive Lasso method can be applied directly in lag selection part. In order to focus on the main idea of Lasso-type cointegration rank selection, theoretical results for fixed dimensional VECM are derived in this chapter. Monto Carlo simulations show that the

Lasso-type method works well for VECM determination up to 16 dimensions, which is beyond the ability of Johansen's method.

In Chapter 3, I extend the Lasso-type method for VECM determination to high dimensions, where the dimension is allowed to increase with the number of observations. Such an extension of VECM requires completely different statistical treatment: For the first, consistency of covariance matrix estimation depends on the ratio between dimension and number of observations; for the second, many standard results on convergence in distribution, e.g. central limit theorem, can't be applied in this setting. Although there are literatures dealing with high dimensional stationary time series, how to deal with the non-stationary components in high dimensions is still an open question, which is the most difficult part in this chapter. To overcome this challenge, I derive the high dimensional strong invariance principle, which approximates the partial sum of stationary time series by Brownian motion. Another difference from the fixed dimensional case is that adaptive Lasso doesn't necessarily lead to variable selection consistency in high dimensions without extra assumptions. This chapter is the first in the literature that derives the conditions under which adaptive Lasso is consistent in variable selection consistency with large probability. Last but not least, only very weak assumptions on the error terms are imposed so that this method is applicable to real financial data. Detailed numerical simulations confirm the power of this method. In the last section, this method is applied to estimate the VECM of European CDS data. The forecast error variance decomposition results show that the countries in the south, i.e. Italy, Spain, Greece, and European banks form two clusters. Within each cluster, the elements are strongly interconnected over time.

Chapter 4 considers an ultra-high dimensional Cointegrated Vector Autoregressive Models (CVAR) where both the loading matrix and the cointegrating vectors are assumed to be sparse. In such case, a step of pre-screening is introduced to avoid overfitting in further steps. After reducing the dimensions significantly, I consider the reduced rank regression approach (RRR) rather than the Lasso-type approach as in Chapters 2 and 3. The consecutive ratios of the eigenvalues from RRR indicate the cointegration rank and the right eigenspace spanned by the first r eigenvectors from RRR approximates the subspace spanned by the cointegrating vectors. The advantage of RRR estimator is that the estimator for the cointegration vectors does no longer suffer from endogeneity bias but the consistency result allows rather small cardinality of the pre-selected subset. This new approach combining pre-screening and RRR has strong practical implications for financial arbitrage strategies.

In Chapter 5, an empirical study with high dimensional high frequency trading data is conducted in order to measure market impacts in a robust way. The main contribution of this chapter is that not only prices but also volumes at different levels from the limit order books across stocks are included in the large dynamic system. This can only be estimated with the appropriate regularized estimation strategy for the underlying high dimensional vector autoregressive model. Different from the traditional approach for fixed dimensional case, here I employ a bootstrapping method to derive directed impulse response function and forecast error variance decomposition. Empirically asymmetric market impacts are identified during the period of Brexit and some stocks in financial industry are significantly leading the prices of others.

Chapters 2-4 are joint work with Melanie Schienle. Chapter 3 is in the third round revision for potential publication at the Journal of Econometrics. Chapter 5, which has been submitted to Review of Financial Studies, is joint work with Melanie Schienle, Shi Chen and Wolfgang Härdle.

2 Lasso-Techniques for Model Determination & Estimation of Non-Stationary Time Series in Higher Dimensions

2.1 Introduction

Complex financial systems are dynamic, multi-dimensional and often contain a large number of non-stationary potentially cointegrated components. Many financial data show such properties, e.g. Credit Default Spreads (CDS), foreign exchange rates, interest rates, etc. The level prices are nonstationary but there exist some linear vectors such that a stationary process can be generated by taking the inner product of the coefficient vector and the price vector. Interestingly, the unobservable stationary process is very important in predicting the return process. Such an economic result was modeled by the vector error correction model (VECM) as introduced in Engle and Granger (1987). But how to determine the number of such linear vectors in subspace sense (or cointegration rank) and how to estimate the whole model become a challenging econometric problem after that. Sequential Johansen test proposed (Johansen, 1988, 1991, 1995) have been the most popular approaching in estimating VECM in the past decades. Other than that, bootstrap based modifications of sequential likelihood tests as e.g. in Cavaliere et al. (2012) or by information criteria such as Chao and Phillips (1999), Wang and Bessler (2005) are proposed in estimating VECM for model selection consistency. However, even for settings greater than dimension two, existing econometric techniques often fail to provide accurate, testable and computationally tractable estimates. As the dimension of variables becomes larger in model financial market, there is thus a need for econometric procedures which can consistently directly determine and estimate appropriate full-dimensional VECM specifications in such higher but still fixed dimensional settings.

In this paper, we provide a Lasso-type technique for consistent and numerically efficient model selection which is feasible for both, standard low but also higher dimensions. Moreover, the proposed adaptive shrinkage method allows for model choice and direct estimation in the same step. Model determination is treated as a joint selection problem of cointegrating rank and VAR lags. Even for moderate cross-section dimensions, the amount of possible combinations of cointegration relations and VAR lags becomes quite large. In this case, we exploit that from a large fixed number of potential cointegration relations, in practice, only a few of them are actually prevalent for the system. In the same way, a small number of VAR lags are considered sufficient for a parsimonious model specification, but within this maximum lag range, our model selection technique is independent from the lag ordering. In this sense, the problem is assumed to be sparse. We show consistency of the variable selection by the proposed Lasso-VECM estimator and derive its asymptotic properties for inference. For refined estimation in particular in larger dimensional finite samples, we provide a refined estimation strategy and derive its statistical properties. Moreover, with only linear computational complexity, all procedures remain computationally tractable also for higher dimensions. A simulation study shows the effectiveness of the proposed techniques in finite samples. This is also illustrated by the empirical example with sovereign CDS data.

Our work builds on the excessive literature of VECM as summarized e.g. in Lütkepohl (2007) as well as on results for Lasso techniques in the standard *i.i.d.* case. Lasso was proposed by Tibshirani (1996) and its asymptotic properties were first studied in Knight and Fu (2000). The adaptive Lasso by Zou (2006) improved on the selection properties by penalizing different variables differently. Yuan and Lin (2006) introduce group-Lasso which allows for simultaneous exclusion and inclusion of certain variables. For the Lasso optimization, there are several standard solution algorithms such as the coordinate descent (Friedman et al., 2007; Friedman et al., 2010, or others), the interior point method (Koh et al., 2007), or the orthant-wise limited-memory Quasi-Newton optimizer (Andrew and Gao, 2007).

But the application of Lasso-type technique to non-stationary time series is new in the literature. There exist some empirical and simulation work employing penalizing algorithms for VECM without mathematical proofs, see e.g. Signoretto and Suykens (2012), Wilms and Croux (2016). While Liao and Phillips (2015) provides a solution of applying Lasso to VECM with theoretical properties, penalizing the eigenvalues of an asymmetric matrix makes its approach different from standard Lasso-type techniques and thus the technical proof challenging.

The rest of the paper is organized as follows. In Section 3.2.1, we derive the Lasso objective function in a VECM specification in order to jointly determine the cointegration rank and the VAR lags. The asymptotic results for model selection and estimation of the proposed Lasso-VECM estimator are stated in Section 2.3. Besides, Section 2.3 also discusses the refinement of the estimate given the estimated rank and lag. Section 2.4 extends the previous econometric analysis to a non *i.i.d.* setting. In Section 3.5 we study the finite-sample performance of the method in several simulation experiments. We also provide an empirical application to sovereign CDS data in Section 3.6. Section 3.7 concludes. All proofs are contained in the Appendix.

2.2 Model and Estimation

In order to illustrate the main ideas of the proposed Lasso methodology, we first derive our Lasso objective function in a simple setting for a known fixed VAR with one lag. Thus model determination here only consists of cointegrating rank selection and estimation. We denote this setup as special case described in Subsection 2.2.1. Results are of independent interest, as such models are widely used in the applied literature. In Subsection 3.3, we generalize the setting to a general unknown VAR with unknown general lag order which then also enters the model selection problem. Thus complete model specification then amounts to both rank and lag order determination.

Throughout the paper, we use the following notation. For $a \in \mathbb{R}^m$, we write $||a||_A^2 = a'Aa$ for any non-singular positive definite matrix A. The corresponding empirical norm is denoted by $||a||_{\tilde{A}}^2 = a'\tilde{A}a$ with a consistent pre-estimate $\tilde{A}$ of A. $||a||_2^2$ denotes the squared l_2 norm. For matrices we use the Frobenius norm $|| \cdot ||_F$.

In general, we consider an m-dimensional $I(1)$ time series Y_t, i.e. Y_t is nonstationary and $\Delta Y_t = Y_t - Y_{t-1}$ is stationary for $t = 1, \dots, T$. Our setup is higher-dimensional, thus m can be large but fixed. Thus obtained results provide a strong improvement of conventional model selection techniques in the VECM setting, but are different from high-dimensional statistical techniques where the dimension can also grow with sample size T.

2.2.1 Special case

For simplicity in this subsection, we assume that Y_t is generated from a VAR(1) process

$$Y_t = A_1 Y_{t-1} + u_t \tag{2.1}$$

with equivalent VECM representation

$$\Delta Y_t \quad = \quad \Pi Y_{t-1} + u_t \tag{2.2}$$

for $t = 1, \dots, T$, where $\Pi = A_1 - I_m$ is an $m \times m$ matrix of rank r with $0 \leqslant r < m$ marking the number of cointegration relations in the system. Π can be decomposed as $\Pi = \alpha\beta'$, where β marks the r long-run cointegrating relations and α is a loading matrix of rank r. This decomposition is unique up to a nonsingular matrix H, so only the space of cointegration relations is identified but not β. In this setting, VECM determination reduces to selection of the correct cointegration rank.

For the error term u_t, we first employ a standard white noise assumption which allows to focus on the key aspects of our Lasso selection procedure while keeping technical results simple.

Assumption 2.2.1. *The error term u_t is i.i.d. distributed with $\mathcal{N}(0, \Sigma_u)$ where Σ_u is a symmetric, positive definite $m \times m$ matrix.*

In Section 2.4, we show how Assumption 2.2.1 can be generalized admitting linear forms of weak dependence. Such a general setting requires changes in the Lasso procedure and leads to different statistical properties of the modified technique.

Our shrinkage selection procedure is based on an available consistent pre-estimate of the cointegration matrix Π. It is well known, that for the model setting in (2.2) and Assumption 2.2.1 the standard least squares estimator

$$\widetilde{\Pi} = \Big(\sum_{t=1}^{T} \Delta Y_t Y_{t-1}'\Big)\Big(\sum_{t=1}^{T} Y_{t-1} Y_{t-1}'\Big)^{-1} \tag{2.3}$$

is consistent while its asymptotic properties depend on the unknown cointegration rank. Moreover, the least squares estimate $\widetilde{\Sigma}_u = \frac{1}{T}\sum_{t=1}^{T}(\Delta Y_t - \widetilde{\Pi} Y_{t-1})(\Delta Y_t - \widetilde{\Pi} Y_{t-1})'$ of the error variance-covariance matrix Σ_u is also consistent (see e.g. Lütkepohl, 2007).

The distribution of $\widetilde{\Pi}$ relies on a Q-transformation of Y_t, which allows to disentangle stationary and nonstationary components. It pre-multiplies all elements in (2.2) from the left with the specific matrix Q defined as follows

$$Q = \begin{bmatrix} \beta' \\ \alpha_\perp' \end{bmatrix} \qquad Q^{-1} = \begin{bmatrix} \alpha(\beta'\alpha)^{-1} & \beta_\perp(\alpha_\perp'\beta_\perp)^{-1} \end{bmatrix}$$

where $\alpha_\perp$ and $\beta_\perp$ denote the orthogonal complement of α and β respectively.[1] Note in particular, that the $I(1)$ assumption on Y_t ensures that $\beta'\alpha$ and $\alpha_\perp'\beta_\perp$ are non-singular component matrices in $r \times r$ and $(m-r) \times (m-r)$ respectively, thus appearing inverses in Q^{-1} exist and all matrices are well-defined.

Thus by Q-transformation, we obtain a new vector $Z_t = QY_t = [(\beta' Y_t)', (\alpha_\perp' Y_t)']' = [Z_{1,t}', Z_{2,t}']'$ decomposed into a distinct stationary and nonstationary part. In particular by definition, the first component $Z_{1,t}$ of dimension r is stationary and the $(m-r)$-dimensional remainder $Z_{2,t}$ is a unit root process.

For determining the cointegration rank, we therefore aim at empirically disentangling the stationary part $Z_{1,t}$ from the non-stationary $Z_{2,t}$ with the help of a Lasso-type procedure. The basic principle of standard Lasso-type methods is to determine the number of covariates in a linear model according to a penalized loss-function

[1] For $m \geqslant r$, we denote by $M_\perp$ an orthogonal complement of the $m \times r$ matrix M with $rk(M) = r$. Thus $M_\perp$ is any $m \times (m-r)$ matrix with $rk(M_\perp) = m - r$ and $M'M_\perp = 0$.

criterion. Likewise, the determination of the cointegration rank in (2.2) amounts to distinguishing the vectors spanning the cointegration space from the basis of its orthogonal complement. This is equivalent to separating the non-zero singular values of Π from the zero ones, where the number of non-zero singular values corresponds to the rank. Thus, the corresponding loading matrix for $\beta' Y_{t-1}$ is α while the remainder $\beta'_\perp Y_{t-1}$ should get loading zero. We say the underlying model has a sparse structure with respect to the rank if $m/r = c_1$ and $c_1 \gg 1$. In this case, which we consider as practically prevalent in the higher-dimensional setting, only a very limited number r of cointegration relationships occur while there are potentially many options m. The problem is more sparse, the larger c_1. In such cases, Lasso-type methods are tailored to detecting corresponding non-zero loadings.

To construct a Lasso-type objective function for rank selection, we require a pre-estimate for β and $\beta_\perp$ respectively, which we obtain from the QR decomposition (with column-pivoting)[2] of $\widetilde{\Pi}'$ as

$$\begin{aligned} \widetilde{\Pi} &= \widetilde{R}'\widetilde{S}' \\ &= \begin{bmatrix} \widetilde{R}'_{1,m\times r} & \widetilde{R}'_{2,m\times(m-r)} \end{bmatrix} \begin{bmatrix} \widetilde{S}'_{1,r\times m} \\ \widetilde{S}'_{2,(m-r)\times m} \end{bmatrix} \end{aligned} \tag{2.4}$$

where $\widetilde{S}$ is an orthonormal matrix, i.e. $\widetilde{S}'\widetilde{S} = I$. $\widetilde{R}$ is an upper triangular matrix [3]and further properties of this decomposition can be found in Stewart (1984). Column-pivoting orders columns in R according to size putting zero-columns at the end.[4] Since $\widetilde{\Pi}$ is a matrix of full-rank and also a consistent estimate of Π, the lower diagonal elements of the last $(m-r)$ columns of the matrix $\widetilde{R}'$ are expected to be small, converging to zero asymptotically at unit root speed T. This is shown in Lemma 2.2.1, where we derive convergence results of the QR-decomposition components $\widetilde{R}$ and $\widetilde{S}$ from the least squares pre-estimate $\widetilde{\Pi}'$.

Lemma 2.2.1. *Let Assumption 2.2.1 hold for $\widetilde{\Pi}$ in* (2.3)*. We denote by $\widetilde{R}'_1$ the first r and by $\widetilde{R}'_2$ the last $m-r$ columns of $\widetilde{R}'$ in the QR-decomposition* (3.7) *of $\widetilde{\Pi}'$. Let β be orthonormal and H be some $(r \times r)$-orthonormal matrix. Then*

[2] To avoid confusion between the orthogonal matrix Q from QR-decomposition and the Q matrix defined previously, we write the former as matrix S.

[3] Such a decomposition exists for any real squared matrix. It is unique for invertible $\widetilde{\Pi}$ if all diagonal entries of $\widetilde{R}$ are fixed to be positive. There are several numerical algorithms like Gram-Schmidt or the Householder reflection which yield the numerical decomposition.

[4] Generally, column pivoting uses a permutation on R such that its final elements $R(i,j)$ fulfill: $|R(1,1)| \geqslant |R(2,2)| \geqslant \ldots \geqslant |R(m,m)|$ and $R(k,k)^2 \geqslant \sum_{i=k+1}^{j} R(i,j)^2$.

$$\begin{aligned} \|\widetilde{S}_1 - \beta H\|_F &= O_p(\frac{1}{T}) \\ \|\widetilde{R}_2\|_F &= O_p(\frac{1}{T}) \\ \sqrt{T} vec(\widetilde{R}_1' H - \alpha) &\rightarrow_d N(0, \Sigma_{z1z1}^{-1} \otimes \Sigma_u) \end{aligned}$$

where $\frac{1}{T}\sum_{t=1}^{T} \beta' Y_{t-1} Y_{t-1}' \beta \rightarrow_p \Sigma_{z1z1}$. *More rigorously,* $\frac{1}{\sqrt{T}}\sum_{t=1}^{[Ts]} \beta' Y_{t-1} \rightarrow_p B_{z1}(s)$ *and* Σ_{z1z1} *is the covariance matrix of Brownian motion* $B_{z1}(s)$.

Remark 2.2.1. 1. *An orthonormal version of* β *in* $\Pi' = \beta\alpha'$ *can always be constructed for the cointegration space e.g. by using the Gram-Schmidt algorithm. It is unique up to rotation, i.e. up to any orthonormal matrix* H. *The form of* Σ_{z1z1} *depends on the specific representation of* β *in the cointegration space.*

2. *If we directly have* $\Pi' = \beta\alpha'$ *as QR-decomposition, i.e.* $H = I_r$. *Then*

$$\sqrt{T} vec(\widetilde{R}_1' - \alpha)_{\mathcal{A}_{QR}} \rightarrow_d N(0, (\Sigma_{z1z1}^{-1} \otimes \Sigma_u)_{\mathcal{A}_{QR}})$$

where $\mathcal{A}_{QR} = \{(i + (j-1)m) \cdot \mathbb{I}_{(i \geqslant j)}, i = 1, 2, \ldots, m, j = 1, 2, \ldots, r\}$. *The subscript* $\mathcal{A}$ *notation denotes for a vector* v *a subvector of elements* v_i *with* $i \in \mathcal{A}$, *and for a matrix* A *a submatrix containing only elements* $a_{i,j}$ *with* $i \in \mathcal{A}, j \in \mathcal{A}$.

Lemma 2.2.1 clearly shows that the last $m-r$ columns of $\widetilde{R}'$ converge to zero at rate T, faster than the $\sqrt{T}$- rate of the first r stationary columns. We exploit this idea in constructing adaptive weights for a model selection consistent Lasso procedure, which put a faster diverging penalty on true zero singular values of Π and less on the non-zero ones corresponding to the underlying stationary components. In particular, we expect that zero columns in R' can be easily detected by adaptive Lasso, as non-zero columns estimated as close to zero would converge slower than true zero components approach zero according to Lemma 2.2.1. Therefore relating penalties in adaptive Lasso to inverses of these initial estimates shrinks true zero components faster to zero than the other ones, which results in a higher penalty for the true zero parts and the detection of the appropriate basis for the cointegration space. Then this adaptive penalty causes the number of non-zero columns in the penalized estimate of R' to produce a consistent estimate for the rank r of Π. Hence elements $\widehat{R}(i,j)$ of $\widehat{R}$ minimize the following criterion over all $R(i,j)$ for $i, j = 1, \ldots, m$

$$\sum_{t=1}^{T} \| \Delta Y_t - R' \widetilde{S}' Y_{t-1} \|_{\Sigma_u^{-1}}^2 + \sum_{i,j=1}^{m} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^{\gamma}} |R(i,j)| \qquad (2.5)$$

where $\widetilde{R}(i,j)$ is the (i,j)th element of an un-penalized pre-estimate $\widetilde{R}$ generated from the QR-decomposition of $\widetilde{\Pi}'$ in (3.7). The penalization parameter λ and the weight γ for adaptiveness are fixed and in practice pre-determined in a data-driven way. See the simulation and application in Sections 3.5 and 3.6 for details. We then obtain an estimate of the true cointegration rank $\hat{r}$ from (2.5) as $\hat{r} = \text{rank}(\hat{R})$, where $\text{rank}(\hat{R})$ equals the number of non-zero columns in $\hat{R}'$. Another advantage of such an objective function is that even non-zero columns in $\hat{R}'$ can still have zero elements, which exploits the sparsity structure of R sufficiently and thus leads to extra efficient gains. We use a GLS-type loss function in (2.5) and also in the subsequent subsection for efficiency purposes in general cases of Σ_u. All are operationalized with the corresponding FGLS approach by minimizing the corresponding empirical norm with pre-estimated $\widetilde{\Sigma}_u^{-1}$.

Due to the properties of the QR-decomposition with column-pivoting, non-zero columns of R' still have many zero elements which is also reflected by the estimates obtained from the adaptive Lasso procedure above. This is different from two-step estimates obtained from sequential likelihood pre-tests or information criteria. In a higher-dimensional setting, however, this case might be prevalent as for any given cointegration relationship, there might be a substantial number of variables which remain unaffected by it. Such type of efficiency gain from sparsity is impossible if the penalty was directly imposed e.g. on the eigenvalues of Π, compare Liao and Phillips (2015).

Moreover, compared with existing literature for our setting, our approach features several advantages: Firstly, the employed QR-decomposition is real-valued without further constraints on the matrix $\widetilde{\Pi}$. Thus the Lasso criterion (2.5) only contains real-valued elements and can be minimized with standard optimization techniques. In higher-dimensions, however, a corresponding eigenvalue decomposition would most likely contain complex values leading to a non-standard harmonic function optimization problem in a respective Lasso objective function. Secondly, the form of the Lasso objective functions is linear in coefficients and therefore straightforward to implement relying on available numerically efficient standard Lasso procedures. So our method is direct and ready to use. And thirdly, the form of our objective function directly allows to employ sparsity constrains for efficient estimation which seems important in particular in higher dimensional settings.

2.2.2 General case

Now we generalize the special case to settings with a general unknown VAR structure. Suppose the general structure of the true process $\{Y_t\}$ is

$$\Delta Y_t \quad = \quad \Pi Y_{t-1} + B_1 \Delta Y_{t-1} + \cdots + B_P \Delta Y_{t-P} + u_t \tag{2.6}$$

for $t = 1, \ldots, T$. As before, the dimension of $\{Y_t\}$ is m, the rank of Π is $r < m$. We set the maximum possible lag length P as sufficiently large but fixed independent of T, such that it is an upper bound for the true lag p, i.e. $p < P$. In this case,

$B_{p+1}, \dots, B_P$ are all zero matrices. For sparsity, additionally $P/p = c_2$ with $c_2 \gg 1$.

The econometric analysis of VECM in the general case also relies on the decomposition of a transformed Y_t into a stationary and a non-stationary component. Its existence is generally guaranteed by the Granger representation theorem (see Engle and Granger (1987)) which requires the following assumptions:

Assumption 2.2.2. *1. The roots for $|(1-z)I_m - \Pi z - \sum_{j=1}^{p} B_j(1-z)z^j| = 0$ is either $|z| = 1$ or $|z| > 1$.*

2. The number of roots lying on the unit circle is $m - r$.

3. The matrix $\alpha_\perp'(I_m - \sum_{i=1}^{p} B_i)\beta_\perp$ is nonsingular.

Note that in the special case, these assumptions are trivially met by the chosen setup.

For estimation purposes, we rewrite the general VECM defined in (3.13), for $t = 1, \dots, T$, in matrix notation as

$$\Delta Y = \Pi Y_{-1} + B\Delta X + U \tag{2.7}$$

where $\Delta Y = [\Delta Y_1, \dots, \Delta Y_T]$, $Y_{-1} = [Y_0, \dots, Y_{T-1}]$, $B = [B_1, \dots, B_P]$, $\Delta X = [\Delta X_0, \dots, \Delta X_{T-1}]$ where $\Delta X_{t-1} = \left[\Delta Y_{t-1}', \dots, \Delta Y_{t-P}'\right]'$ and $U = [u_1, \dots, u_T]$. W.l.o.g, $Y_k = 0$ for $k \leqslant 0$. Moreover, we denote $\Gamma_t = [Y_{t-1}'\beta, \Delta Y_{t-1}', \dots, \Delta Y_{t-P}']'$. Under Assumptions 2.2.1 and 3.2.2, it holds by Lemma 1 in Toda and Phillips (1993)

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{[Ts]} \Gamma_t \to_p B_\Gamma(s) \tag{2.8}$$

where $B_\Gamma(s)$ is a Brownian motion with covariance

$$\Sigma_{\Gamma\Gamma} = \begin{pmatrix} \Sigma_{z1z1} & \Sigma_{z1\Delta x} \\ \Sigma_{\Delta xz1} & \Sigma_{\Delta x\Delta x} \end{pmatrix} \tag{2.9}$$

The least squares estimate for (3.2) is denoted by $[\widetilde{\Pi}_{ls}, \widetilde{B}_{ls}]$, which will be used to get the consistent estimate $\widetilde{\Sigma}_u = \frac{1}{T-mP+1}(\Delta Y - \widetilde{\Pi}_{ls}Y_{-1} - \widetilde{B}_{ls}\Delta X)(\Delta Y - \widetilde{\Pi}_{ls}Y_{-1} - \widetilde{B}_{ls}\Delta X)'$ of Σ_u (see e.g. Lütkepohl, 2007).

For model selection, we disentangle the joint lag-rank selection problem by employing a Frisch-Waugh-idea in the VECM model (3.2). With this, we obtain two independent criteria for lag and rank choice which can be computed separately. For rank selection, the partial least squares pre-estimate $\widetilde{\Pi}$ can be obtained from the corresponding partial model when removing the effect of ΔX in ΔY and Y_{-1} by regressing ΔYM on $Y_{-1}M$ with $M = I_T - \Delta X'(\Delta X \Delta X')^{-1}\Delta X$. Thus it is

$$\widetilde{\Pi} = \Big(\Delta YMY'_{-1}\Big)\Big(Y_{-1}MY'_{-1}\Big)^{-1} \tag{2.10}$$

Lemma 2.2.2. *Under Assumptions 2.2.1 and 3.2.2, the partial least squares estimate $\widetilde{\Pi}$ defined in (3.4) satisfies*

$$\begin{array}{ll} & vec[Q(\widetilde{\Pi} - \Pi)Q^{-1}D_T] \\ \rightarrow_d & \left[\begin{array}{l} N(0, \Sigma^{-1}_{z1z1.\Delta x} \otimes \Sigma_v) \\ vec\Big\{\Sigma_v^{1/2}(\int_0^1 W^{\dagger}_{m-r} dW'_m)'(\int_0^1 W^{\dagger}_{m-r} W^{\dagger\prime}_{m-r} ds)^{-1}(\alpha'_{\perp}\Sigma_u\alpha_{\perp})^{-\frac{1}{2}}\Theta^{-1}_{22} \end{array} \right] \end{array}$$

where $D_T = diag(\sqrt{T}I_r, TI_{m-r})$, $\Sigma_v = Q\Sigma_u Q'$, $Z_{-1} = \beta' Y_{-1}$, $\frac{1}{T}Z_{-1}MZ'_{-1} \rightarrow_p \Sigma_{z1z1.\Delta x} = \Sigma_{z1z1} - \Sigma_{z1\Delta x}\Sigma^{-1}_{\Delta x \Delta x}\Sigma_{\Delta x z1}$ with all the compoment covariance matrices defined in (2.9)*;*

$W^{\dagger}_{m-r} = (\alpha'_{\perp}\Sigma_u\alpha_{\perp})^{-\frac{1}{2}}[0_{(m-r)\times r}, I_{m-r}]\Sigma_v^{\frac{1}{2}} W_m$, and $W^{\dagger}_{m-r}, W_m$ are standard Brownian motions with dimension $m-r, m$ respectively and the exact from of Θ is defined as (2.20) *and* (2.21) *in the proof.*

Here we have $\Sigma_{z1z1.\Delta x}$ instead of Σ_{z1z1} in the variance part of the stationary component due to the partial estimation problem and the residual maker M. In the non-stationary component, the term Θ appears due to the lagged differenced term ΔX.

Lemma 2.2.2 shows that $\widetilde{\Pi}$ is a consistent estimate. Thus we can employ the idea of the previous subsection for rank selection and separate the problem into stationary and nonstationary parts as in the special case. We thus obtain for the components of the QR-decomposition $\widetilde{\Pi} = \widetilde{R}'\widetilde{S}'$:

Lemma 2.2.3. *Let Assumptions 2.2.1 and 3.2.2 hold for $\widetilde{\Pi}$ in* (3.4)*. We denote by $\widetilde{R}'_1$ the first r and by $\widetilde{R}'_2$ the last $m-r$ columns of $\widetilde{R}'$ in the QR-decomposition* (3.7) *of $\widetilde{\Pi}'$ defined in* (3.4)*. Let β be orthonormal and H be a $(r \times r)$-orthonormal matrix.*

$$\begin{array}{rcl} ||\widetilde{S}_1 - \beta H||_F & = & O_p(\dfrac{1}{T}) \\ ||\widetilde{R}_2||_F & = & O_p(\dfrac{1}{T}) \\ \sqrt{T}vec(\widetilde{R}'_1 H - \alpha) & \rightarrow_d & N(0, \Sigma^{-1}_{z1z1.\Delta x} \otimes \Sigma_u) \end{array}$$

where $\frac{1}{T}\beta' Y_{-1}MY'_{-1}\beta \rightarrow_p \Sigma_{z1z1.\Delta x}$ and $\Sigma_{z1z1.\Delta x}$ is defined as in Lemma 2.2.2.

Remark 2.2.2. *Discussions analogous to Remark 2.2.1 also apply here.*

Thus from Lemma 2.2.2 and 2.2.3, we can construct a corresponding adaptive Lasso procedure as an analogue to (2.5) in vector form. Hence components $\hat{R}(i,j)$ of $\hat{R}$ minimize the following criterion over all $R(i,j)$ for $i,j = 1,\dots,m$

$$\|vec(\Delta YM) - (MY_{-1}'\widetilde{S} \otimes I_m)vec(R')\|^2_{I_T\otimes\Sigma_u^{-1}} + \sum_{i,j=1}^{m} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^\gamma}|R(i,j)| \tag{2.11}$$

where now $\widetilde{R}(i,j)$ is from the QR-decomposition of $\widetilde{\Pi}'$ in the partial model (3.4). We choose the cointegration rank as $\hat{r} = \text{rank}(\hat{R})$, where $\text{rank}(\hat{R})$ is the number of non-zero columns in $\hat{R}'$.

Likewise, for independent lag selection, the effect of the nonstationary term Y_{-1} in (3.2) must be filtered out in ΔY and ΔX for unbiased estimation in the partial model via regression of ΔYC on ΔXC with $C = I_T - Y_{-1}'(Y_{-1}Y_{-1}')^{-1}Y_{-1}$. Thus we obtain $\hat{B}$ as minimizing the following objective function over all components $B_k(i,j)$ for $k = 1,\dots,P$ and $i,j = 1,\dots,m$

$$||vec(\Delta YC) - (C\Delta X' \otimes I_m)vec(B)||^2_{I_T\otimes\Sigma_u^{-1}} + \sum_{k=1}^{P}\sum_{i,j=1}^{m} \frac{\lambda_{i,j,T}^{lag,k}}{|\check{B}_k(i,j)|^\gamma}|B_k(i,j)| \tag{2.12}$$

for fixed tuning parameters $\lambda_{i,j,T}^{lag,k}, \gamma$, where γ here and in the rank selection (2.11) might differ. Moreover, the pre-estimate $\check{B}$ in the adaptive Lasso weight can be taken from the partial least squares estimate $\widetilde{B} = (\Delta YC\Delta X')(\Delta XC\Delta X')^{-1}$ due to consistency. Though in practice, especially with larger dimensions and lags, multicollinearity effects in ΔX are quite likely to occur which cause the least squares estimate to become numerically instable. Therefore we also consider a robust ridge type pre-estimate $\widetilde{B}^R$ as $\check{B}$, which can be obtained from

$$\begin{aligned}\widetilde{B}^R \quad &= \arg\min \|vec(\Delta YC) - (C\Delta X' \otimes I_m)vec(B)\|^2 \\ &+\nu_T \textstyle\sum_{k=1}^{P}\sum_{i,j=1}^{m} |B_k(i,j)|^2\end{aligned} \tag{2.13}$$

The following Theorem 2.2.1 shows that this pre-estimate is also consistent for appropriate choices of tuning parameters

Theorem 2.2.1. *If the tuning parameter ν_T in the ridge regression (2.13) satisfies $\frac{\nu_T}{\sqrt{T}} \to_p 0$, then $\sqrt{T}(\widetilde{B}^R - B) = O_p(1)$ under Assumptions 2.2.1 and 3.2.2.*

As in the case of rank selection, a lag k should be included into the model, whenever $\hat{B}_k$ from the Lasso selection (2.12) is different from zero. Thus, in contrast to other model selection criteria, a Lasso-type procedure allows for the inclusion of non-consecutive lags, which we consider an additional advantage of the procedure. We obtain an estimate $\hat{p}$ of the true lag length from (2.12) as $\hat{p} = \max_{1\leqslant k\leqslant P}\{k|\hat{B}_k \neq 0\}$.

Remark 2.2.3. *1. The residual transformation C in the lag selection criterion is similar to the second term of the PIC statistics introduced in Chao and*

Phillips (1999). Moreover, the lag selection procedure is independent of the unknown rank.

2. *Ridge regression can be further extended to elastic net (see Zou and Hastie (2005)) or sure independence screening (see Fan and Lv (2008)) for a sparse, consistent and numerically stable pre-estimate.*

3. *The separate two-step approach for rank and lag length can help alleviate the numerical instability caused by multi-collinearity. The following subsection will show that a larger than necessary lag P has no effect on model selection consistency which is the main focus of the paper. Only obtained estimates of β suffer from a corresponding efficiency loss which can be cured with a refinement (see Subsection 2.3.3).*

2.3 Main Results for Model Selection Consistency

In this section, we state the asymptotic properties of the adaptive Lasso-VECM procedure for the special and the general cases.

2.3.1 Model Selection Consistency for special VECM

The following theorem derives the statistical properties of the estimate from our adaptive Lasso procedure (2.5) for special VECM.

Theorem 2.3.1. *Suppose that $\lambda^{rank}_{i,j,T}/\sqrt{T} \to 0$ and $T^{\frac{1}{2}(\gamma-1)}\lambda^{rank}_{i,j,T} \to \infty$ and $\Pi' = SR$ is a QR decomposition with column pivot. Then under Assumptions 2.2.1 the objective function (2.5) yields:*

1. $\lim_{T\to\infty} \mathbb{P}(\mathcal{A}^*_T = \mathcal{A}) = 1$
   where $\mathcal{A}$ is the set of indices for the non-zero elements of $vec(R')$, $\mathcal{A}^*_T$ is the set of indices for the non-zero elements of $vec(\hat{R}'_T)$ derived in (2.5).

2. $\sqrt{T}vec(\hat{R}'_T - R')_{\mathcal{A}} \to_d \mathrm{N}(0, (\Sigma_{z1z1} \otimes \Sigma_u^{-1})^{-1}_{\mathcal{A}}(\Sigma_{z1z1} \otimes \Sigma_u^{-1})_{\mathcal{A}}(\Sigma_{z1z1} \otimes \Sigma_u^{-1})^{-1}_{\mathcal{A}})$ *if* $r > 0$

Theorem 2.3.1 shows that our method is consistent in variable selection i.e., it chooses the right rank and the correct sparse pattern with probability one. This is our primary and main concern. Note that the weight function in the adaptive Lasso procedure is crucial to achieve this property.

Additionally, the second part of the theorem gives the asymptotic distribution of the adaptive Lasso-VECM estimate. It is asymptotically unbiased converging to a normal distribution at the standard stationary speed $\sqrt{T}$. The complicated structure of the variance matrix is due to the sparse structure of R in our Lasso procedure. However, this estimate suffers from endogeneity bias caused by the naive estimate of β. More specifically, the bias in estimating β from the least squares estimate $\widetilde{\Pi}$ depends on the term $\int_0^1 dW_m(s)W_m(s)'$, in which the integrand $W_m(s)$ and the differential part $dW_m(s)$ are the same Brownian motion, thus dependent. The bias could be further decreased if had the form $\int_0^1 dW_1(s)W_2(s)'$ with W_1, W_2 independent Brownian motions. The latter can be achieved by reduced rank regression, see Anderson (2002) for detailed asymptotics. Therefore, it is recommended to update β after obtaining a consistent estimate for r. For details we refer to Subsection 2.3.3.

Remark 2.3.1. *If we treat each column of R' as a group and apply adaptive group Lasso, with a similar proof we can show that the right rank can still be estimated consistently. The asymptotic distribution of the nonzero columns of R' is the same as Lütkepohl (2007), page 277, where the true β is assumed to be known. Though, the sparse structure is neglected which would produce inferior finite-sample estimation and prediction results when the dimension increases.*

2.3.2 Model Selection Consistency for general VECM

First, we show the result for the cointegrating rank selection according to criterion (2.11) which uses the residual transformation M in order to focus on the respective partial effect in the general VECM. The structure of the result resembles the one of the special case.

Theorem 2.3.2. *Suppose that $\lambda^{rank}_{i,j,T}/\sqrt{T} \to 0$ and $T^{\frac{1}{2}(\gamma-1)}\lambda^{rank}_{i,j,T} \to \infty$. Under Assumptions 2.2.1 and 3.2.2 with the same notation for $\mathcal{A}$ as in Theorem 2.3.1 the objective function (2.11) yields*

1. $\lim_{T\to\infty} \mathbb{P}(\mathcal{A}^*_T = \mathcal{A}) = 1$
   where $\mathcal{A}^*_T$ is index set of the non-zero elements of vec$(\hat{R}')$ in (2.11).
2. $\sqrt{T}vec(\hat{R}'_T - R')_{\mathcal{A}} \to_d \mathrm{N}(0, (\Sigma_{z1z1.\Delta x}\otimes\Sigma_u^{-1})^{-1}_{\mathcal{A}}(\Sigma_{z1z1.\Delta x}\otimes\Sigma_u^{-1})_{\mathcal{A}}(\Sigma_{z1z1.\Delta x}\otimes\Sigma_u^{-1})^{-1}_{\mathcal{A}})$
   for $r > 0$.

Thus Theorem 2.3.2 yields rank selection consistency. Moreover, for the variance of the estimates of the non-zero components in R, a smaller P closer to the true p would provide additional efficiency gains. Using valid restrictions on irrelevant components of ΔX_{t-1} variation in $\Sigma_{z1z1.\Delta x}$ could be reduced. As our focus here is on model selection, however, this is a secondary concern and we point to Subsection 2.3.3 for refined estimation.

In addition to the rank, for general VECM, we also need to determine the correct lag in a separate procedure. The following theorem shows the results using the Lasso lag selection criterion (2.12) with adaptive weights from a ridge regression pre-estimate $\widetilde{B}^R$. In this way, we account for prevalent multicollinearity effects in particular in settings with higher dimensions and large lag lengths.

Theorem 2.3.3. *Suppose that* $\lambda_{i,j,T}^{lag,k}/\sqrt{T} \to 0$ *and* $T^{\frac{1}{2}(\gamma-1)}\lambda_{i,j,T}^{lag,k} \to \infty$*. Then the lag objective function (2.12) yields:*

1. $\lim_{T\to\infty} \mathbb{P}(\mathcal{B}_T^* = \mathcal{B}) = 1$;
   where $\mathcal{B}$ *is the set of indices for the non-zero elements of* $vec(B)$, $\mathcal{B}_T^*$ *is the set of indices for the non-zero elements of* $vec(\hat{B})$ *in* (2.12)

2. $\sqrt{T}vec(\hat{B}_T'-B')_{\mathcal{B}} \to_d \mathrm{N}(0, (\Sigma_{\Delta x\Delta x.z1}\otimes\Sigma_u^{-1})_{\mathcal{B}}^{-1}(\Sigma_{\Delta x\Delta x.z1}\otimes\Sigma_u^{-1})_{\mathcal{B}}(\Sigma_{\Delta x\Delta x.z1}\otimes\Sigma_u^{-1})_{\mathcal{B}}^{-1})$
   where $\Sigma_{\Delta x\Delta x.z1} = \Sigma_{\Delta x\Delta x} - \Sigma_{\Delta xz1}\Sigma_{z1z1}^{-1}\Sigma_{z1\Delta x}$ *with all the compoment covariance matrices defined in* (2.9).

Thus lag selection is consistent i.e., the true lags are selected with probability 1 even if they are non-consecutive. For estimation of the coefficients in the relevant lag components, as in the case for the rank, we find asymptotic normality and unbiasedness at the standard stationary speed. Different to the rank selection result in Theorem 2.3.2, however, the variance component $\Sigma_{\Delta x\Delta x.z1}$ only depends on the true rank r automatically and a pre-estimate for it is not necessary. This results from the different speed of convergence which asymptotically separates the stationary cointegrated component $Z_{1,t-1}$ and the nonstationary parts. In this sense, penalized estimates of lag coefficients are more efficient than the ones for R.

2.3.3 Refined Model Estimation in Higher Dimensions

With our proposed adaptive Lasso techniques, we can select the true model with probability one for sufficiently many observations. Although both model selection criteria (2.11) and (2.12) also yield consistent estimates for the coefficients of appropriate variables, there is, however, substantial room for improvement on the estimation side in particular in finite samples for higher dimensions. For pure model estimation in higher dimensions, we therefore suggest a refined procedure for α and B_k with $k \in \{1, \ldots, p\}$ which is still of Lasso type but no longer adaptive. With a focus on model estimation, given the pre-selected rank and lag, we propose a pure Lasso procedure rather than an adaptive variant. While the latter is targeted at consistent model selection, a pure Lasso estimate performs better in estimation and prediction (see Bühlmann and Van De Geer (2011) for the comparison of different variants of Lasso).

Besides, we use an improved estimate $\tilde{\beta}^\dagger$ of β from reduced rank regression (see Ahn and Reinsel (1990) and Anderson (2002)), which does not suffer from endogeneity bias and yields improved finite sample performance. Please note, that generally $\tilde{\beta}^\dagger$ an efficient estimate of $\tilde{\beta}^\dagger$ relies on a precise estimate for the rank by matrix perturbation theory, as well as a consistent estimate for the lag p. Therefore in particular in higher-dimensional sparse settings, it can only be employed in the estimation refinement step and is no option for the pre-step in model selection.

We thus obtain estimates $\hat{\hat{\alpha}}, \hat{\hat{B}}_1, \ldots, \hat{\hat{B}}_p$ as minimizers of

$$\begin{aligned} &\sum_{t=1}^{T} ||\Delta Y_t - \alpha \tilde{\beta}^{\dagger\prime} Y_{t-1} - \sum_{k=1}^{p} B_k \Delta Y_{t-k}||^2_{\Sigma_u^{-1}} \\ &+ \sum_{i=1}^{m} \sum_{j=1}^{r} \lambda_{i,j,T}^{rank} |\alpha(i,j)| + \sum_{k=1}^{p} \sum_{i,j=1}^{m} \lambda_{i,j,T}^{lag,k} |B_k(i,j)| \end{aligned} \tag{2.14}$$

where $\lambda_{i,j,T}^{rank}, \lambda_{i,j,T}^{lag,k}$ are tuning parameters. For no penalty $\lambda_{i,j,T}^{rank} = \lambda_{i,j,T}^{lag,k} = 0$, we recover the reduced rank regression estimates for α and B^p from (2.14).

We show that with appropriate choices of tuning parameters, the penalized estimates from (2.14) are consistent and yield the same asymptotic variance as the ones from reduced rank regression, while its solution is sparse in finite samples and thus improves the mean squared error in general. Though as the simulations in Section 3.5 will confirm, their finite-sample performance, however, is superior in particular for estimation but also for prediction.

Theorem 2.3.4. *Denote $B^p = [B_1, \ldots, B_p]$. If $\lambda_{i,j,T}^{rank}/\sqrt{T} \to_p 0$ and $\lambda_{i,j,T}^{lag,k}/\sqrt{T} \to_p 0$, then the solution to problem (2.14) under Assumptions 2.2.1 and 3.2.2 satisfies:*

$$\sqrt{T}\Big(vec([\hat{\hat{\alpha}}_T, \hat{\hat{B}}_T^p]) - vec([\alpha, B^p]) \Big) \sim_d N(0, \Sigma_{\Gamma^p\Gamma^p}^{-1} \otimes \Sigma_u)$$

where $\Gamma_t^p = [Y_{t-1}'\beta, \Delta Y_{t-1}', \ldots, \Delta Y_{t-p}']'$ and $\frac{1}{T}\sum_{t=1}^{T} \Gamma_t^p \Gamma_t^{p\prime} \to_p \Sigma_{\Gamma^p\Gamma^p}$.

Theorem 2.3.4 shows that asymptotically, the penalized estimate has the same distribution as the reduced rank estimate. This is in contrast to the adaptive estimates in Theorem 2.3.2 and 2.3.3. In finite samples, however, the variances of nonzero Lasso estimates are smaller than those from the reduced rank because variables with small coefficients are excluded from the model, see Section 3.5 for details. Thus even if Lasso estimates may suffer from finite-sample bias, the overall mean squared error might still be superior. Secondly, although reduced rank estimates are consistent, i.e. in finite samples, estimates of irrelevant zero components are small but might add up influencing estimation and prediction significantly. The advantage of the penalized estimate in higher dimensions might result from the fact that the assumption of sparsity in α and B_j becomes increasingly justified with dimensions more than 3, i.e. often only a small group of leading variables has impact on the whole system while many others are irrelevant for the rest. Besides, the tuning parameter can be chosen in the same manner as in univariate case.

2.4 Extension to Dependent Error Terms

In this section we illustrate how Assumption 2.2.1 on *i.i.d.* innovations can be relaxed. Generally, independent error terms help to simplify the theoretical analysis but for real data they are often hard to justify. Therefore we provide explicit results for more general weak dependence structures and show in which way they effect and deteriorate estimates for α and β. We illustrate the main effects in the setting of the special case only.

Assumption 2.4.1. *In the special VECM as* (2.2) *the error term can admit the following linear dependence structure*

$$u_t = \sum_{j=0}^{\infty} \kappa_j w_{t-j} \quad \textit{with} \sum_{j=0}^{\infty} j\|\kappa_j\|_2 < \infty.$$

where $w_t \overset{iid}{\sim} \mathcal{N}(0, \Sigma_w)$ *and* Σ_w *is positive definite matrix.*

Assumption 2.4.1 is stronger than absolute summability due to the convergence of unit root processes.

Lemma 2.4.1. *Under Assumption 2.4.1, the least squares estimate for* Π *in* (2.2) *is biased and satisfies*

$$Q(\widetilde{\Pi} - \Pi)Q^{-1} \xrightarrow{P} [Q\Upsilon\Sigma_{z1z1}^{-1}, 0_{m\times(m-r)}]$$

For the exact form of Υ *as well as the asymptotic distribution of* $\widetilde{\Pi}$ *we refer to the Appendix (see Lemma 2.A.2).*

The term Υ measures the correlation between u_t and $Z_{1,t-1}$ due to the autocorrelation of u_t under Assumption 2.4.1.

Define $\Xi = \begin{bmatrix} \beta' \\ \beta'_\perp \end{bmatrix}$ as in the proof for Lemma 2.2.1, we have

$$\Xi(\widetilde{\Pi}' - \Pi' - \beta\Sigma_{z1z1}^{-1}\Upsilon') = \Xi\widetilde{\Pi}' - \begin{bmatrix} \alpha' + \Sigma_{z1z1}^{-1}\Upsilon' \\ 0 \end{bmatrix}$$

By a similar argument as for Lemma 2.2.1, we can conclude that

Lemma 2.4.2. *By the same notation as in Lemma 2.2.1 and under Assumption*

2.4.1, the following results hold:

$$\begin{aligned} ||\widetilde{S}_1 - \beta H||_F &= O_p(\frac{1}{T}) \\ ||\widetilde{R}_2||_F &= O_p(\frac{1}{T}) \\ \sqrt{T} vec(\widetilde{R}_1' H - \alpha - \Upsilon \Sigma_{z1z1}^{-1}) &\rightarrow_d N(0, \Sigma_{z1z1}^{-1} \otimes \Sigma_w) \end{aligned}$$

Due to the bias term, we can't expect that the selection result from (2.5) is consistent element-wise, but consistency in rank could still hold when the penalty term is modified. The estimate $\hat{R}$ is obtained by minimizing the follwing objective function row-wise in $R(i,)$ for $i = 1, \ldots, m$

$$\sum_{t=1}^{T} \parallel \Delta Y_t - R' \widetilde{S}' Y_{t-1} \parallel_2^2 + \sum_{i=1}^{m} \frac{\lambda_{i,T}^{rank}}{||\widetilde{R}(i,)||_2^{\gamma}} ||R(i,)||_2 \qquad (2.15)$$

Different from (2.5), we penalize each row in R as a group, similar to Yuan and Lin (2006), Wang and Leng (2008). Therefore, there could be zero and non-zero rows in $\hat{R}$, but non-zero rows have no zero elements. By Lemma 2.4.2, the penalty on the first r rows of R would be bounded and the penalty on the last $m - r$ rows explodes. Thus consistency of the estimate from (2.15) in rank selection is expected. Besides, the first term in (2.15) is equivalent to the ordinary least squares problem rather than a generalized least squares because we penalize the each row in R as a whole. The statistical property is given in Proposition 2.4.1.

Proposition 2.4.1. *Given Assumption 2.4.1, suppose that* $\lambda_{i,T}^{rank}$ *satisfies* $\frac{\lambda_{i,T}^{rank}}{\sqrt{T}} \rightarrow 0$ *and* $T^{\gamma-1}\lambda_{i,T}^{rank} \rightarrow \infty$*, the solution to* (2.15) *is consistent in selecting the right rank.*

When the dimension is higher, the variance of $\hat{R}$ from (2.5) generally increases due to the non-sparse structure within non-zero rows of $\hat{R}$.

2.5 Simulations

In this section, we investigate the finite-sample performance of the proposed model selection methodology. Moreover, we also study estimation and prediction performance of the refined Lasso estimates in comparison to reduced rank regression. This includes standard settings of dimension three for comparison with existing low dimensional techniques. But in particular, we focus on cases up to dimension eight and sixteen with a thorough simulation study of model selection quality as well as the estimation and forecast fit. Such higher dimensional specifications are not feasible with available standard techniques and provide a substantial generalization to the common bivariate illustrations in this literature.

In all model specifications we consider independent multivariate Gaussian innovations with covariance matrix $\Sigma_u = [\rho^{|i-j|}]_{i,j=1}^{m}$ for two particular cases $\rho = 0.0$ and $\rho = 0.6$. Thus our specifications include cases of strong cross-sectional dependence. The chosen vanishing pattern of correlations corresponds e.g. to increasing geographical distance in the case of the sovereign CDS application in Section 3.6. For these settings, we use the general FGLS-type empirical versions of the objective functions (2.11) and (2.12) for model selection with least squares estimate $\widetilde{\Sigma}_u$ for Σ_u.

For each model, we provide simulation results based on $T = 200$ and $T = 500$ observations corresponding to roughly one year and 2.5 years of working days in financial data. In each setting, simulation and model selection are repeated for $b = 100$ times.

For transparency, we report all results dependent on the choice of tuning parameters γ and λ in the adaptive Lasso procedure. Thus for each setting, we show all results on a two-dimensional grid of $\lambda = cT^{1/2-\varepsilon}$ and γ where $\varepsilon = 0.1$ and c takes all integers from 1 to 3 and γ ranges from 2 to 5 in steps of 1. We use the same penalty λ for both rank and lag selection, which could in practice be refined with different tuning parameters for each criterion. Although lag and rank selection work independently, we found that the efficiency point in lag selection in Theorem 2.3.3 leads to superior finite-sample choices of p, which can then be used in setting P for numerical efficient rank selection in (2.11). In the literature, BIC is a standard way to choose tuning parameters. For comparison, we mark the BIC-selection of (γ, c) in the Tables by underlining respective values. They are obtained as minimizing the following criteria:

$$\begin{aligned} BIC_{rank} &= \log|\Sigma_{res}| + \frac{\log T}{T}\hat{r}(\lambda,\gamma)m \\ BIC_{lag} &= \log|\Sigma_{res}| + \frac{\log T}{T}\hat{p}(\lambda,\gamma)m^2 \end{aligned}$$

The first term of the criteria is the goodness of fit measured by the determinant of the covariance matrix of the residuals, and the second terms are the penalty. Because we are interested in the selection results of how many columns in R' or lags B_k should be kept in the model, the number of free coefficients are $\hat{r}m$ or $\hat{p}m^2$ respectively.

Simulations for model selection are done in `R`. Lasso is implemented with the package `lbfgs` (called through `Rcpp` for faster speed) which can solve the penalized model for a fixed tuning parameter numerically very efficiently. For pure model estimation part, we use the R-package `grpreg`, which works for a sequence of tuning parameters and has the implemented option to select the optimal tuning parameter by BIC.

2.5.1 Model Specifications

For the standard three dimensional case, we choose a setting considered in Chao and Phillips (1999) for comparison purposes. For the higher dimensions, at each level of model complexity with given dimension, cointegration rank and lag length, our simulation settings are randomly chosen from all possible VECM specifications satisfying the Assumption 3.2.2. In particular, all appearing unknown elements are drawn independently from $U[-1.5, 1.5]$. We then work with the first specification which satisfies the standard assumptions. In this paper, we consider the following cases:

model 1:	$m = 3$	$r = 2$	$p = 1$
model 2:	$m = 8$	$r = 4$	$p = 1$
model 3:	$m = 8$	$r = 2$	$p = 2$
model 4:	$m = 16$	$r = 8$	$p = 1$

For model 1, we use the following specification:

$$\Delta Y_t = \alpha\beta' Y_{t-1} + B_1 \Delta Y_{t-1} + u_t \tag{2.16}$$

with

$$\alpha\beta' = \begin{bmatrix} -0.25 & 0 \\ 1.2 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -0.5 \end{bmatrix}$$

and

$$B_1 = \begin{bmatrix} 0.25 & 0 & 0 \\ -1.2 & 0.1 & 0 \\ 0 & -0.5 & 0.25 \end{bmatrix}$$

In all other settings, the exact randomly chosen model specifications are provided in 2.C.

2.5.2 Model Selection Results

For the simple three dimensional model 1, rank and lag selection results are reported in Table 2.1. The results indicate that lag selection performs well independently of the exact choice of tuning parameters with almost perfect results. For rank selection in this simplest case, the penalty term should not be too large i.e. we require $c = 1$ with $\gamma = 2$ for good finite-sample performance.

Model 1 ($T = 200,\ \rho = 0.0$)

	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	$\underline{100}/\underline{95}$	100/100	96/100
$\gamma = 3.0$	98/100	80/100	59/100
$\gamma = 4.0$	80/100	50/100	24/100
$\gamma = 5.0$	57/100	22/99	10/98

Model 1 ($T = 500,\ \rho = 0.0$)

	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	$\underline{100}/\underline{99}$	100/100	100/100
$\gamma = 3.0$	100/100	100/100	99/100
$\gamma = 4.0$	100/100	87/100	62/100
$\gamma = 5.0$	88/100	50/100	20/100

Model 1 ($T = 200,\ \rho = 0.6$)

	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	$\underline{100}/\underline{86}$	100/100	92/100
$\gamma = 3.0$	98/100	80/100	58/100
$\gamma = 4.0$	79/100	48/100	27/100
$\gamma = 5.0$	54/100	27/100	14/100

Model 1 ($T = 500,\ \rho = 0.6$)

	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	$\underline{100}/\underline{81}$	100/99	100/100
$\gamma = 3.0$	100/100	100/100	97/100
$\gamma = 4.0$	98/100	89/100	66/100
$\gamma = 5.0$	89/100	55/100	28/100

Table 2.1. Absolute numbers XX/YY of correct model selections by solving (2.11) and (2.12) for $b = 100$ repetitions of model 1 with $m = 3,\ r = 2,\ p = 1$. For each parameter specification, XX denotes the number of correct rank selections while YY is the number of correct lag length identifications. Underlining marks the choice with tuning parameters selected according to BIC.

Models 2 and 3 are both of dimension $m = 8$, where traditional methods cannot be employed either due to inconsistency in theory or because of numerical inefficiency. The selection results for model 2 with $p = 1$ with $r = 4$ in Table 2.2 demonstrate perfect performance in rank and lag selection generally for a wide range of tuning parameters with $c \geqslant 1$ and $\gamma \geqslant 3$. This also holds even for the most difficult case with $\rho = 0.6$ and $T = 200$, while for all other settings the range of acceptable parameters is even wider. In comparison to the simple model 1, larger tuning parameters are preferred both for rank and lag selection due to the higher complexity of the true model. Note that in all cases, the results are based on a ridge regression pre-estimate (2.13) for the lag choice criterion (2.12) in order to handle multicollinearity effects. Lag selection results based on adaptive weights from least squares pre-estimates perform substantially inferior.[5] The increased lag length $p = 2$ with $r = 2$ poses the challenge in model 3. There, in particular in the case of 200 observations, larger tuning parameters are preferred for rank selection.

[5] Results are not reported here but are available on request.

Model 2 ($m = 8$, $r = 4$, $p = 1$, $T = 200$, $\rho = 0.0$)

	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	99/34	100/72	99/84
$\gamma = 3.0$	100/$\underline{97}$	100/100	100/100
$\gamma = 4.0$	100/100	100/100	100/100
$\gamma = 5.0$	$\underline{100}$/100	100/100	100/100

Model 2 ($m = 8$, $r = 4$, $p = 1$, $T = 500$, $\rho = 0.0$)

	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	100/45	100/81	100/$\underline{90}$
$\gamma = 3.0$	100/100	100/100	100/100
$\gamma = 4.0$	100/100	100/100	100/100
$\gamma = 5.0$	$\underline{100}$/100	100/100	100/100

Model 2 ($m = 8$, $r = 4$, $p = 1$, $T = 200$, $\rho = 0.6$)

	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	92/1	100/14	97/33
$\gamma = 3.0$	100/$\underline{88}$	100/99	98/99
$\gamma = 4.0$	100/100	99/100	99/100
$\gamma = 5.0$	$\underline{100}$/100	99/100	99/100

Model 2 ($m = 8$, $r = 4$, $p = 1$, $T = 500$, $\rho = 0.6$)

	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	99/1	100/7	100/16
$\gamma = 3.0$	100/$\underline{88}$	100/99	100/100
$\gamma = 4.0$	100/100	100/100	100/100
$\gamma = 5.0$	$\underline{100}$/100	100/100	100/100

Model 3 ($m = 8$, $r = 2$, $p = 2$, $T = 200$, $\rho = 0.0$)

	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	63/$\underline{91}$	95/98	100/99
$\gamma = 3.0$	100/100	100/100	100/100
$\gamma = 4.0$	100/94	100/65	100/41
$\gamma = 5.0$	$\underline{100}$/41	100/11	100/1

Model 3 ($m = 8$, $r = 2$, $p = 2$, $T = 500$, $\rho = 0.0$)

	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	$\underline{100}$/$\underline{100}$	100/100	100/100
$\gamma = 3.0$	100/100	100/100	100/100
$\gamma = 4.0$	100/100	100/100	100/100
$\gamma = 5.0$	100/100	100/91	100/68

Model 3 ($m = 8$, $r = 2$, $p = 2$, $T = 200$, $\rho = 0.6$)

	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	35/$\underline{63}$	80/80	90/92
$\gamma = 3.0$	92/100	97/99	99/97
$\gamma = 4.0$	98/90	99/48	98/17
$\gamma = 5.0$	$\underline{99}$/13	99/0	99/0

Model 3 ($m = 8$, $r = 2$, $p = 2$, $T = 500$, $\rho = 0.6$)

	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	95/$\underline{69}$	100/85	100/94
$\gamma = 3.0$	100/100	100/100	100/100
$\gamma = 4.0$	100/100	100/100	100/100
$\gamma = 5.0$	$\underline{100}$/99	100/56	100/26

Table 2.2. Absolute numbers of correct rank/lag selections by solving (2.11) and (2.12) for $b = 100$ repetitions for model 2 and 3 with $m = 8,\ r = 2,\ p = 2$. Reporting style is as in Table 2.1.

Model 4 (T = 500, $\rho = 0.0$)				**Model 4** (T = 500, $\rho = 0.6$)			
	$c = 1$	$c = 2$	$c = 3$		$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	$\underline{69}/\underline{98}$	98/100	100/100	$\gamma = 2.0$	$11/\underline{93}$	58/100	84/100
$\gamma = 3.0$	100/100	78/100	46/100	$\gamma = 3.0$	$\underline{100}/100$	95/100	83/100
$\gamma = 4.0$	49/100	11/100	5/100	$\gamma = 4.0$	77/100	48/100	19/100
$\gamma = 5.0$	9/100	2/100	0/100	$\gamma = 5.0$	28/100	10/100	2/100

Table 2.3. Absolute numbers of correct rank/lag selections by solving (2.11) and (2.12) for $b = 100$ repetitions for model 4 with $m = 16$, $r = 8$, $p = 1$. Reporting style is as in Table 2.1.

For model 4, we consider a nonstationary VAR(2) process like in model 1 but of dimension 16, i.e. $m = 16$, $r = 8$ and $p = 1$. Due to the complexity from the higher dimensionality of the model we only report results for $T = 500$. For well-chosen tuning parameters, both rank and lag selection results are perfect. In particular, $\gamma = 2$ with larger c and $\gamma = 3$ with smaller c are crucial for good performance of rank selection. Given the complexity of the model, however, there is still a range of such admissible tuning parameters which ensures robust performance in application scenarios where tuning parameters must be pre-chosen. As for models 2 and 3, we use a ridge regression estimate for $\check{B}$ in the lag selection criterion (2.12).

Simulations show that the lag selection results are generally better than rank selection results. The reason lies in that rank selection problem is based on a pre-estimated cointegrating space, which adds one more source of finite-sample bias.

2.5.3 Estimation Results

For known true model specifications, we estimate all four models above according to the refined Lasso procedure (2.14) and compare estimation fits and one-step ahead forecasts to reduced rank regression. For the case of model 1, we also illustrate their finite-sample advantage if the model is known to the adaptive Lasso estimates from the model selection procedure. In particular, we use $\hat{\Pi}_{adaptive} = \hat{R}'_r \widetilde{S}'_r$ where $\hat{R}'_r$ comprises the first r columns of the solution to the adaptive Lasso rank selection problem (2.11) and $\widetilde{S}'_r$ consists of the first r rows of the orthonormal matrix defined in (3.7).

We generally only report the most difficult case $\rho = 0.6$. We report pointwise empirical quantiles of squared errors over all simulation iterations for Π , B_k and the $1-$step ahead squared forecast error. In particular, we evaluate $||\hat{\Pi}_\star - \Pi||_2^2$ and the

$T = 200$	25%	50%	75%
$\|\|\hat{\hat{\Pi}}_{lasso} - \Pi\|\|_2^2$	$7.974e^{-4}$	$1.376e^{-3}$	$2.588e^{-3}$
$\|\|\hat{\hat{\Pi}}_{ls} - \Pi\|\|_2^2$	$7.536e^{-4}$	$1.424e^{-3}$	$3.004e^{-3}$
$\|\|\hat{\Pi}_{adaptive} - \Pi\|\|_2^2$	$3.902e^{-3}$	$1.807e^{-2}$	$3.370e^{-2}$
$\|\|\hat{\hat{B}}_{1,lasso} - B_1\|\|_2^2$	$1.606e^{-3}$	$2.759e^{-3}$	$4.206e^{-3}$
$\|\|\hat{\hat{B}}_{1,ls} - B_1\|\|_2^2$	$2.246e^{-3}$	$3.561e^{-3}$	$6.258e^{-3}$
$\|\|\Delta\hat{Y}_{T+1,lasso} - \Delta Y^*_{T+1}\|\|_2^2$	$1.617e^{-2}$	$4.527e^{-2}$	$1.032e^{-1}$
$\|\|\Delta\hat{Y}_{T+1,ls} - \Delta Y^*_{T+1}\|\|_2^2$	$1.818e^{-2}$	$3.928e^{-2}$	$1.062e^{-1}$
$T = 500$	25%	50%	75%
$\|\|\hat{\hat{\Pi}}_{lasso} - \Pi\|\|_2^2$	$3.502e^{-4}$	$5.562e^{-4}$	$9.509e^{-4}$
$\|\|\hat{\hat{\Pi}}_{ls} - \Pi\|\|_2^2$	$3.759e^{-4}$	$6.413e^{-4}$	$1.131e^{-3}$
$\|\|\hat{\Pi}_{adaptive} - \Pi\|\|_2^2$	$1.771e^{-3}$	$1.131e^{-2}$	$2.919e^{-2}$
$\|\|\hat{\hat{B}}_{1,lasso} - B_1\|\|_2^2$	$7.979e^{-4}$	$1.195e^{-3}$	$1.990e^{-3}$
$\|\|\hat{\hat{B}}_{1,ls} - B_1\|\|_2^2$	$9.162e^{-4}$	$1.471e^{-3}$	$2.268e^{-3}$
$\|\|\Delta\hat{Y}_{T+1,lasso} - \Delta Y^*_{T+1}\|\|_2^2$	$1.442e^{-2}$	$2.917e^{-2}$	$5.725e^{-2}$
$\|\|\Delta\hat{Y}_{T+1,ls} - \Delta Y^*_{T+1}\|\|_2^2$	$1.257e^{-2}$	$2.605e^{-2}$	$4.507e^{-2}$

Table 2.4. Comparison of different estimation methods for Model 1

same loss function for B_k, where the norm denotes the squared l_2 norm of $vec(\hat{\Pi}_\star - \Pi)$ divided by m^2, in which $\star$ refers to cases where $\hat{\Pi}$ is estimated by Lasso or least squares. We divide by m in order to ensure comparability of results across different dimensions. $\Delta\hat{Y}_{T+1,\star}$ denotes the 1-step ahead forecast based on method $\star$ and ΔY^*_{T+1} is the forecast based on the true model. Again for comparability the squared l_2 norm is divided by m and the reported forecast error is normalized by $\Sigma_u^{-\frac{1}{2}}$.

The results for model 1 indicate the refined estimation leads to superior results if the true model is selected. Besides, refined Lasso estimates of Π and B_1 are overall better than the least squares (LS). In this simple 3-dimensional model, however, the prediction based on the tailored high-dimensional Lasso procedure is dominated by the one of LS due to the inherent sample bias.

For the more complex model 2 with $m = 8$ and $r = 4$, however, Lasso is substantially superior to LS in both estimation and prediction (see Table 2.5). Similar results are reported in Table 2.6 for model 3 and Table 2.7 for model 4. While in the standard low-dimensional model 1, the advantage of using Lasso is not so significant, we find that the more complicated the model is, the more superior becomes the Lasso in particular in estimation. Moreover, the obtained simulation results confirms the advantage of element-wise penalization on the loading matrix over penalization on

$T = 200$	25%	50%	75%
$\|\|\hat{\hat{\Pi}}_{lasso} - \Pi\|\|_2^2$	$8.293e^{-3}$	$1.339e^{-2}$	$2.068e^{-2}$
$\|\|\hat{\hat{\Pi}}_{ls} - \Pi\|\|_2^2$	$3.569e^{-2}$	$5.100e^{-2}$	$7.193e^{-2}$
$\|\|\hat{\hat{B}}_{1,lasso} - B_1\|\|_2^2$	$4.396e^{-3}$	$8.778e^{-3}$	$1.333e^{-2}$
$\|\|\hat{\hat{B}}_{1,ls} - B_1\|\|_2^2$	$2.964e^{-2}$	$3.946e^{-2}$	$5.289e^{-2}$
$\|\|\Delta\hat{Y}_{T+1,lasso} - \Delta Y^*_{T+1}\|\|_2^2$	2.998	5.872	15.150
$\|\|\Delta\hat{Y}_{T+1,ls} - \Delta Y^*_{T+1}\|\|_2^2$	4.332	10.510	16.390
$T = 500$	25%	50%	75%
$\|\|\hat{\hat{\Pi}}_{lasso} - \Pi\|\|_2^2$	$3.035e^{-3}$	$4.384e^{-3}$	$5.882e^{-3}$
$\|\|\hat{\hat{\Pi}}_{ls} - \Pi\|\|_2^2$	$1.021e^{-3}$	$1.532e^{-2}$	$2.107e^{-2}$
$\|\|\hat{\hat{B}}_{1,lasso} - B_1\|\|_2^2$	$2.302e^{-3}$	$3.537e^{-3}$	$4.676e^{-3}$
$\|\|\hat{\hat{B}}_{1,ls} - B_1\|\|_2^2$	$9.562e^{-3}$	$1.302e^{-2}$	$1.784e^{-2}$
$\|\|\Delta\hat{Y}_{T+1,lasso} - \Delta Y^*_{T+1}\|\|_2^2$	$6.553e^{-1}$	2.279	5.329
$\|\|\Delta\hat{Y}_{T+1,ls} - \Delta Y^*_{T+1}\|\|_2^2$	1.208	2.908	6.604

Table 2.5. Comparison of different estimation methods for Model 2

eigenvalues/singular values only. In the latter case, e.g. Liao and Phillips (2015), the "one-step" approach is not able to take the sparse structure of loading matrix in higher dimension into account.

2.6 Empirical Example

In this section, we illustrate our Lasso-VECM approach on daily sovereign Credit Default Swap (CDS) data. In particular, we are interested in long-term interconnections of sovereign default risk within the European Union, which we identify as cointegration relations between CDS spreads of respective countries. We focus on CDS of five countries: United Kingdom (Y_1), Germany (Y_2), France (Y_3), Belgium (Y_4) and Italy (Y_5). Our analysis is based on data from Bloomberg and contains the period from 01.2013 to 12.2015.

We set the maximum lag length to P=5, and determine the tuning parameters c and γ of the Lasso-VECM procedure via BIC among all c and γ on the grid from 1 to 5 with step 0.5. λ is defined the same as in the simulation. For the BIC choice of

$T = 200$	25%	50%	75%
$\|\|\hat{\hat{\Pi}}_{lasso} - \Pi\|\|_2^2$	$5.365e^{-3}$	$7.092e^{-3}$	$9.005e^{-3}$
$\|\|\hat{\hat{\Pi}}_{ls} - \Pi\|\|_2^2$	$3.655e^{-2}$	$4.578e^{-2}$	$5.861e^{-2}$
$\|\|\hat{\hat{B}}_{1,lasso} - B_1\|\|_2^2$	$2.694e^{-3}$	$3.813e^{-3}$	$4.911e^{-3}$
$\|\|\hat{\hat{B}}_{1,ls} - B_1\|\|_2^2$	$3.809e^{-2}$	$4.769e^{-2}$	$6.229e^{-2}$
$\|\|\hat{\hat{B}}_{2,lasso} - B_2\|\|_2^2$	$1.633e^{-2}$	$1.683e^{-2}$	$1.740e^{-2}$
$\|\|\hat{\hat{B}}_{2,ls} - B_2\|\|_2^2$	$3.183e^{-2}$	$3.183e^{-2}$	$3.720e^{-2}$
$\|\|\Delta\hat{Y}_{T+1,lasso} - \Delta Y^*_{T+1}\|\|_2^2$	$1.467e^{-1}$	$3.232e^{-1}$	$6.040e^{-1}$
$\|\|\Delta\hat{Y}_{T+1,ls} - \Delta Y^*_{T+1}\|\|_2^2$	$5.232e^{-1}$	1.179	2.824
$T = 500$	25%	50%	75%
$\|\|\hat{\hat{\Pi}}_{lasso} - \Pi\|\|_2^2$	$1.939e^{-3}$	$2.357e^{-3}$	$2.888e^{-3}$
$\|\|\hat{\hat{\Pi}}_{ls} - \Pi\|\|_2^2$	$1.175e^{-2}$	$1.641e^{-2}$	$2.248e^{-2}$
$\|\|\hat{\hat{B}}_{1,lasso} - B_1\|\|_2^2$	$1.046e^{-3}$	$1.404e^{-3}$	$1.696e^{-3}$
$\|\|\hat{\hat{B}}_{1,ls} - B_1\|\|_2^2$	$1.329e^{-2}$	$1.741e^{-2}$	$2.318e^{-2}$
$\|\|\hat{\hat{B}}_{2,lasso} - B_2\|\|_2^2$	$1.635e^{-2}$	$1.667e^{-2}$	$1.688e^{-2}$
$\|\|\hat{\hat{B}}_{2,ls} - B_2\|\|_2^2$	$1.909e^{-2}$	$2.197e^{-2}$	$2.343e^{-2}$
$\|\|\Delta\hat{Y}_{T+1,lasso} - \Delta Y^*_{T+1}\|\|_2^2$	$8.695e^{-2}$	$1.481e^{-1}$	$2.495e^{-1}$
$\|\|\Delta\hat{Y}_{T+1,ls} - \Delta Y^*_{T+1}\|\|_2^2$	$2.527e^{-1}$	$5.200e^{-1}$	1.013

Table 2.6. Comparison of different estimation methods for Model 3

	25%	50%	75%
$\|\|\hat{\hat{\Pi}}_{lasso} - \Pi\|\|_2^2$	$5.654e^{-2}$	$6.065e^{-2}$	$6.540e^{-2}$
$\|\|\hat{\hat{\Pi}}_{ls} - \Pi\|\|_2^2$	$9.650e^{-2}$	$1.159e^{-1}$	$1.374e^{-1}$
$\|\|\hat{\hat{B}}_{1,lasso} - B_1\|\|_2^2$	$1.718e^{-2}$	$2.032e^{-2}$	$2.374e^{-2}$
$\|\|\hat{\hat{B}}_{1,ls} - B_1\|\|_2^2$	$8.274e^{-2}$	$1.004e^{-1}$	$1.185e^{-2}$
$\|\|\Delta\hat{Y}_{T+1,lasso} - \Delta Y^*_{T+1}\|\|_2^2$	7.623	17.190	39.280
$\|\|\Delta\hat{Y}_{T+1,ls} - \Delta Y^*_{T+1}\|\|_2^2$	16.940	33.020	61.280
	25%	50%	75%
$\|\|\hat{\hat{\Pi}}_{lasso} - \Pi\|\|_2^2$	$5.297e^{-2}$	$5.506e^{-2}$	$5.859e^{-2}$
$\|\|\hat{\hat{\Pi}}_{ls} - \Pi\|\|_2^2$	$7.435e^{-2}$	$8.232e^{-2}$	$9.599e^{-2}$
$\|\|\hat{\hat{B}}_{1,lasso} - B_1\|\|_2^2$	$2.223e^{-2}$	$2.381e^{-2}$	$2.519e^{-2}$
$\|\|\hat{\hat{B}}_{1,ls} - B_1\|\|_2^2$	$5.705e^{-2}$	$6.479e^{-2}$	$7.428e^{-2}$
$\|\|\Delta\hat{Y}_{T+1,lasso} - \Delta Y^*_{T+1}\|\|_2^2$	7.078	12.900	26.600
$\|\|\Delta\hat{Y}_{T+1,ls} - \Delta Y^*_{T+1}\|\|_2^2$	9.052	17.210	36.290

Table 2.7. Comparison of different estimation methods for Model 4

$c = 1$ and $\gamma = 1.5$, we obtain the fitted model with rank 1 and lag 1,

$$\hat{\Pi} = \begin{bmatrix} 1.9500 & -2.5220 & 1.3145 & 0.6722 & -1.2419 \\ -0.5994 & 0.7754 & -0.4042 & -0.2067 & 0.3818 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{bmatrix} \tag{2.17}$$

$$\hat{B}_1 = \begin{bmatrix} -1.7930 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & -1.4120 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & -0.1833 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & -1.3662 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{bmatrix} \tag{2.18}$$

$\hat{\Pi}$ in (2.17) in shows that there is only one (stable) cointegration relationship during the Euro-crisis periods. Accordingly, only UK and Germany, the leading economies in EU, are affected by the long-run cointegration relations. In the short run, France and Belgium act quite independently and the CDS spreads for France show more persistence. Notably, the data of Italy behaves like a random walk, which is consistent with the intuition that the risk is totally unpredictable due to the insolvency.

2.7 Conclusion

In this paper, we propose a new method to determine the number of cointegration relationships and VAR lags in a sparse vector error correction model. We derive the asymptotic properties of the Lasso-VECM estimator. Our method is computationally efficient and easily applicable in higher dimensions. As for standard Lasso techniques, future work needs to address theoretically and practically optimal ways for a data-driven choice of the tuning parameters. Also, statistical properties of post-adaptive Lasso selection estimates and predictions require a thorough investigation which is beyond the scope of the present paper. Moreover, we also work on extending the theoretical results to the high-dimensional case with r and p growing with the sample size T. Statistical techniques for this case, however, are fundamentally different and demanding. They will be investigated in a separate technical paper.

2.A Proofs

Proof of Lemma 2.2.1

Define $D_T = diag(\sqrt{T}I_r, TI_{m-r})$ and $E_Q = Q(\widetilde{\Pi} - \Pi)Q^{-1}D_T$. Asymptotically each element in matrix E_Q is finite in probability.
Define an orthonormal matrix $\Xi = \begin{bmatrix} \beta' \\ \beta'_\perp \end{bmatrix}$, then $\Xi\Pi' = \begin{bmatrix} \alpha' \\ 0 \end{bmatrix}$, and

$$\begin{aligned} \Xi\widetilde{\Pi}' &= \Xi\Pi' + \Xi Q' D_T^{-1} E_Q' Q^{-1\prime} \\ &= \begin{bmatrix} \alpha' \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{1}{\sqrt{T}} I_r & \frac{1}{T}\beta'\alpha_\perp \\ 0 & \frac{1}{T}\beta'_\perp\alpha_\perp \end{bmatrix} E_Q' Q^{-1\prime} \end{aligned} \tag{2.19}$$

From the last equality, we know that the $m-r$ smallest eigenvalues of $\Xi\widetilde{\Pi}'$ are of small order of $\frac{1}{T}$ in probability, i.e., $O_p(\frac{1}{T})$. The QR-decomposition of $\widetilde{\Pi}' = \widetilde{S}\widetilde{R}$ where $\widetilde{R}$ is an upper triangular matrix. Define

$$\widetilde{R} = \begin{bmatrix} \widetilde{R}_{11} & \widetilde{R}_{12} \\ 0 & \widetilde{R}_{22} \end{bmatrix}$$

Therefore, by the properties of QR-decomposition with pivoting, the rank of $\Xi\widetilde{\Pi}'$ converges to r asymptotically is equivalent to that $\widetilde{R}_{22}$ is negligible. Since $\widetilde{R}_{22}$ is an upper-triangular matrix, the smallest $m-r$ eigenvalues of $\Xi\widetilde{\Pi}'$ converge to zero at the same rate as the diagonal elements of $\widetilde{R}_{22}$. Due to the properties of column pivoting, all the elements in $\widetilde{R}_{22}$ have order $O_p(\frac{1}{T})$. Besides, all the diagonal elements in $\widetilde{R}_{11}$ are significantly different from zero otherwise the asymptotic rank of $\Xi\widetilde{\Pi}'$ is smaller than r.

$$\begin{aligned}\widetilde{\Pi}' &= \widetilde{S}\widetilde{R} \\ &= \begin{bmatrix} \widetilde{S}_1 & \widetilde{S}_2 \end{bmatrix} \begin{bmatrix} \widetilde{R}_1 \\ \widetilde{R}_2 \end{bmatrix}\end{aligned}$$

Thus

$$\Xi\widetilde{\Pi}' = \begin{bmatrix} \beta'\widetilde{S}_1\widetilde{R}_1 + \beta'\widetilde{S}_2\widetilde{R}_2 \\ \beta'_\perp\widetilde{S}_1\widetilde{R}_1 + \beta'_\perp\widetilde{S}_2\widetilde{R}_2 \end{bmatrix}$$

By the last equality of (2.19), $\beta'_\perp\widetilde{S}_1\widetilde{R}_1 + \beta'_\perp\widetilde{S}_2\widetilde{R}_2$ satisfies $O_p(\frac{1}{T})$ element-wise. Thus we conclude each element in $\beta'_\perp\widetilde{S}_1$ has order $O_p(\frac{1}{T})$ due to the fact that $\widetilde{R}_1$ has full row rank and $\widetilde{S}_1$ is orthogonal to $\widetilde{S}_2$. Thus $||\beta'_\perp\widetilde{S}_1||_F = O_p(\frac{1}{T})$, which means that the subspace generated by $\widetilde{S}_1$ is a consistent estimate for that generated by β. Moreover, since $\Xi\widetilde{S}_1$ is an orthonormal matrix, $\beta'\widetilde{S}_1$ converges to an orthonormal one, denoted by H at the rate of T. Mathematically, by

$$I_r = (\beta'\widetilde{S}_1)'(\beta'\widetilde{S}_1) + (\beta'_\perp\widetilde{S}_2)'(\beta'_\perp\widetilde{S}_2)$$

and

$$||\beta'_\perp\widetilde{S}_1||_F = O_p(\frac{1}{T})$$

the following can be derived

$$\|I_r - (\beta'\widetilde{S}_1)'(\beta'\widetilde{S}_1)\|_F^2 = O_p(\frac{1}{T^2})$$

or equivalently in finite dimensional case,

$$\|I_r - (\beta'\widetilde{S}_1)'(\beta'\widetilde{S}_1)\|_2 = O_p(\frac{1}{T})$$

which means that all eigenvalues of $\beta'\widetilde{S}_1$ converge to unit circle from inside at rate of T. Equivalently, for some orthonormal matrix H, it holds that

$$\beta'(\widetilde{S}_1 - \beta H) = O_p(\frac{1}{T})$$

If $\|\beta'(\widetilde{S}_1 - \beta H)\|_F$ converges to zero asymptotically, we have either $(\widetilde{S}_1 - \beta H) \in S(\beta_\perp)$ or $\|\widetilde{S}_1 - \beta H\|_F \to 0$. The first possiblity is excluded by $||\beta'_\perp\widetilde{S}_1||_F \to 0$, Therefore, we conclude that asymptotically, $\widetilde{S}_1$ and β characterize the same space with equivalent

matrix representations.
Lastly, due to the faster rate of convergence for β, the asymptotic distribution of the estimate for α is not affected by the finite-sample error. By the sparse structure of $\widetilde{R}$ imposed by QR-decomposition, the asymptotic distribution depends on the relevant part only, which is similar to that of adaptive Lasso.

□

Lemma 2.A.1. *With the notation defined in Section 3.3, we have*

$$\begin{aligned} \frac{1}{T}\Delta X C \Delta X' &\to_p \Sigma_{\Delta x \Delta x.z1} \\ \frac{1}{\sqrt{T}} vec(UC\Delta X') &\to_p N(0, \Sigma_{\Delta x \Delta x.z1} \otimes \Sigma_u) \\ \frac{1}{T} UCU' &\to_p \Sigma_u \end{aligned}$$

where $\Sigma_{\Delta x \Delta x.z1} = \Sigma_{\Delta x \Delta x} - \Sigma_{\Delta x z1}\Sigma_{z1z1}^{-1}\Sigma_{z1\Delta x}$.

$$\begin{aligned} & \frac{1}{T}\Delta X C \Delta X' \\ = \ & \frac{1}{T}\sum_{t=1}^{T} \Delta X_{t-1}\Delta X'_{t-1} - \frac{1}{T}\Delta X Y'_{-1}(Y_{-1}Y'_{-1})^{-1}Y_{-1}\Delta X' \\ = \ & \frac{1}{T}\sum_{t=1}^{T} \Delta X_{t-1}\Delta X'_{t-1} \\ - \ & \tfrac{1}{T}[\tfrac{1}{\sqrt{T}}\textstyle\sum_{t=1}^{T} \Delta X_{t-1}Z'_{1,t-1}, \tfrac{1}{T}\textstyle\sum_{t=1}^{T} \Delta X_{t-1}Z'_{2,t-1}]\begin{pmatrix} \frac{1}{T}Z_{1,-1}Z'_{1,-1} & \frac{1}{T^{3/2}}Z_{1,-1}Z'_{2,t-1} \\ \frac{1}{T^{3/2}}Z_{2,t-1}Z'_{1,-1} & \frac{1}{T^2}Z_{2,t-1}Z'_{2,t-1} \end{pmatrix}^{-1}\begin{bmatrix} \frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_{1,t-1}\Delta X'_{t-1} \\ \frac{1}{T}\sum_{t=1}^{T} Z_{2,t-1}\Delta X'_{t-1} \end{bmatrix} \\ = \ & \frac{1}{T}\sum_{t=1}^{T} \Delta X_{t-1}\Delta X'_{t-1} \\ - \ & [\tfrac{1}{T}\textstyle\sum_{t=1}^{T} \Delta X_{t-1}Z'_{1,t-1}, \tfrac{1}{T^{3/2}}\textstyle\sum_{t=1}^{T} \Delta X_{t-1}Z'_{2,t-1}]\begin{pmatrix} \frac{1}{T}Z_{1,-1}Z'_{1,-1} & \frac{1}{T^{3/2}}Z_{1,-1}Z'_{2,t-1} \\ \frac{1}{T^{3/2}}Z_{2,t-1}Z'_{1,-1} & \frac{1}{T^2}Z_{2,t-1}Z'_{2,t-1} \end{pmatrix}^{-1}\begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T} Z_{1,t-1}\Delta X'_{t-1} \\ \frac{1}{T^{3/2}}\sum_{t=1}^{T} Z_{2,t-1}\Delta X'_{t-1} \end{bmatrix} \end{aligned}$$

Because $\frac{1}{T}\sum_{t=1}^{T} \Delta X_{t-1}Z'_{1,t-1} \to_p \Sigma_{\Delta xz1}$, $\frac{1}{T^{3/2}}\sum_{t=1}^{T} \Delta X_{t-1}Z'_{2,t-1} \to_p 0$. Thus the first result follows.

The second claim follows naturally because we have already proved the covariance matrix of ΔXC.

$$\begin{aligned}
& \frac{1}{T}UCU \\
= & \frac{1}{T}\sum_{t=1}^{T} u_t u_t' - \frac{1}{T}UY_{-1}'(Y_{-1}Y_{-1}')^{-1}Y_{-1}U' \\
= & \frac{1}{T}\sum_{t=1}^{T} u_t u_t' \\
- & \tfrac{1}{T}[\tfrac{1}{\sqrt{T}}\textstyle\sum_{t=1}^{T} u_t Z_{1,t-1}', \tfrac{1}{T}\textstyle\sum_{t=1}^{T} u_t Z_{2,t-1}'] \begin{pmatrix} \frac{1}{T}Z_{1,-1}Z_{1,-1}' & \frac{1}{T^{3/2}}Z_{1,-1}Z_{2,t-1}' \\ \frac{1}{T^{3/2}}Z_{2,t-1}Z_{1,-1}' & \frac{1}{T^2}Z_{2,t-1}Z_{2,t-1}' \end{pmatrix}^{-1} \begin{bmatrix} \frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_{1,t-1}u_t' \\ \frac{1}{T}\sum_{t=1}^{T} Z_{2,t-1}u_t' \end{bmatrix} \\
= & \frac{1}{T}\sum_{t=1}^{T} u_t u_t' + O_p(\frac{1}{T}) \rightarrow_p \Sigma_u
\end{aligned}$$

□

Proof for Lemma 2.2.2

By the same argument as that for the special case, we have

$$\begin{aligned}
& Q(\widetilde{\Pi} - \Pi)Q^{-1}D_T \\
= & QUMY_{-1}'Q'D_T^{-1}(D_T^{-1}QY_{-1}MY_T^{-1}Q'D_T^{-1})^{-1} \\
= & QUMZ_{-1}'D_T^{-1}(D_T^{-1}Z_{-1}MZ_{-1}'D_T^{-1})^{-1}
\end{aligned}$$

where $Z_{-1}' = [Z_{1,-1}', Z_{2,t-1}']$ and $Z_{1,-1}', Z_{2,t-1}'$ satisfy the following process:

$$\begin{aligned}
\Delta Z_{1,-1}M &= \beta'\alpha Z_{1,-1}M + \beta'\xi \\
Z_{2,-1}M &= Z_{2,-1}M + \alpha_\perp'\xi
\end{aligned}$$

where $\xi = U - U\Delta X'(\Delta X\Delta X')^{-1}\Delta X$.
In order to derive the asymptotic distributions, we also need some notations as follows: By pre-multiply all the terms of general VECM by Q:

$$\Delta Y_t = \Pi Y_{t-1} + B\Delta X_{t-1} + u_t$$

We have

$$\Delta Z_t = Q\Pi Q^{-1}Z_{t-1} + \psi_t \tag{2.20}$$

where $\psi_t = QB\Delta X_{t-1} + v_t$, $v_t = Qu_t$ with covariance matrix Σ_v and

$$\psi_t = \Theta(L)v_t \tag{2.21}$$

Define $\Theta = \Theta(1)$ and Θ_{22} as the bottom-right $(m-r)\times(m-r)$ submatrix of Θ.

1. Distribution of Error Terms:
According to Ahn and Reinsel (1990), $\frac{1}{\sqrt{T}}U\Delta X' = O_p(1)$, $\frac{1}{T}\Delta X\Delta X' = O_p(1)$ and $\frac{1}{\sqrt{T}}\Delta X_{t-1} = O_p(\frac{1}{\sqrt{T}})$. Therefore we have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{[Ts]}\xi_t \Rightarrow_d \Sigma_u^{\frac{1}{2}}W_m(s)$$

since $\frac{1}{T}\sum_{t=1}^{T}\Delta X_{t-1} \to_p 0$.

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\xi_t\xi_t' &= \frac{1}{T}UU' - \frac{1}{T}U\Delta X(\Delta X\Delta X')^{-1}\Delta XU' \\ &= \frac{1}{T}UU' - \frac{1}{T}(\frac{1}{\sqrt{T}}U\Delta X)(\frac{1}{T}\Delta X\Delta X')^{-1}(\frac{1}{\sqrt{T}}\Delta XU') \\ &\to_p \Sigma_u\end{aligned}$$

2. Distribution of $D_T^{-1}Z_{-1}MZ_{-1}'D_T^{-1}$:

$$D_T^{-1}Z_{-1}MZ_{-1}'D_T^{-1} = \begin{bmatrix} \frac{1}{T}Z_{1,-1}MZ_{1,-1}' & \frac{1}{T^{3/2}}Z_{1,-1}MZ_{2,-1}' \\ \frac{1}{T^{3/2}}Z_{2,-1}MZ_{1,-1}' & \frac{1}{T^2}Z_{1,-1}MZ_{2,-1}' \end{bmatrix}$$

The distributions of each block in the matrix would be analyzed as follows:

$$\begin{aligned}\frac{1}{T}Z_{1,-1}MZ_{1,-1}' &= \frac{1}{T}Z_{1,-1}Z_{1,-1}' - \frac{1}{T}Z_{1,-1}\Delta X'(\Delta X\Delta X)^{-1}\Delta XZ_{1,-1}' \\ &= \frac{1}{T}Z_{1,-1}Z_{1,-1}' - \frac{1}{T}Z_{1,-1}\Delta X'(\frac{1}{T}\Delta X\Delta X)^{-1}\frac{1}{T}\Delta XZ_{1,-1}' \\ &\to_p \Sigma_{z1z1} - \Sigma_{z1\Delta x}\Sigma_{\Delta x\Delta x}^{-1}\Sigma_{\Delta xz1}\end{aligned}$$

$$\begin{aligned}\frac{1}{T^{3/2}}Z_{1,-1}MZ_{2,-1}' &= \frac{1}{T^{3/2}}Z_{1,-1}Z_{2,-1}' - \frac{1}{T^{3/2}}Z_{1,-1}\Delta X'(\Delta X\Delta X)^{-1}\Delta XZ_{2,-1}' \\ &= \frac{1}{T^{3/2}}Z_{1,-1}Z_{2,-1}' - \frac{1}{T^{3/2}}Z_{1,-1}\Delta X'(\frac{1}{T}\Delta X\Delta X)^{-1}\frac{1}{T}\Delta XZ_{2,-1}'\end{aligned}$$

By the result from Ahn and Reinsel (1990), $\frac{1}{T}\Delta X Z'_{2,-1} = O_p(1)$, $\frac{1}{T}Z_{1,-1}\Delta X' = O_p(1)$ and $\frac{1}{T}Z_{1,-1}Z'_{2,-1} = O_p(1)$. Therefore, the blocks on upper-right and bottom-left converge to zero in probablity to zero.

$$\begin{aligned}\frac{1}{T^2}Z_{2,-1}MZ'_{2,-1} &= \frac{1}{T^2}Z_{2,-1}Z'_{2,-1} - \frac{1}{T}\frac{1}{T}Z_{2,-1}\Delta X'(\frac{1}{T}\Delta X\Delta X)^{-1}\frac{1}{T}\Delta X Z'_{2,-1} \\ &\rightarrow_d \Theta_{22}(\alpha'_\perp\Sigma_u\alpha_\perp)^{1/2}\int_0^1 W_{m-r}(s)W'_{m-r}(s)ds(\alpha'_\perp\Sigma_u\alpha_\perp)^{1/2}\Theta'_{22}\end{aligned}$$

3. Distribution of $QUMZ'_{-1}D_T^{-1}$:

$$\begin{aligned}QUMZ'_{-1}D_T^{-1} &= [\frac{1}{\sqrt{T}}VMZ_{1,-1}, \frac{1}{T}VMZ_{2,-1}] \\ &- [\frac{1}{\sqrt{T}}V\Delta X'(\frac{1}{T}\Delta X\Delta X')\frac{1}{T}\Delta X Z_{1,-1}, \frac{1}{\sqrt{T}}V\Delta X'(\frac{1}{T}\Delta X\Delta X')\frac{1}{T^{\frac{3}{2}}}\Delta X Z_{2,-1}] \\ &= [\frac{1}{\sqrt{T}}VMZ_{1,-1}, \frac{1}{T}VZ_{2,-1} + \rho_p(1)]\end{aligned}$$

The last equality follows from $\frac{1}{T^{\frac{3}{2}}}\Delta X Z_{2,-1} \rightarrow_p 0$ as shown in Ahn and Reinsel (1990). Since we have shown that $\frac{1}{T}Z_{1,-1}MZ'_{1,-1} \rightarrow_p \Sigma_{z1z1.\Delta x}$, $\frac{1}{\sqrt{T}}vec(VMZ_{1,-1}) \rightarrow_d N(0, \Sigma_{z1z1.\Delta x} \otimes \Sigma_v)$. Besides, the $\frac{1}{T}VZ_{2,-1}$ converges in distribution to

$$\Sigma_v^{\frac{1}{2}}[\int_0^1 W_{m-r}(s)dW_m(s)']'(\alpha'_\perp\Sigma_u\alpha_\perp)^{1/2}\Theta'_{22}$$

To derive the desired result, we just need to combine all the separate terms.

□

Proof of Lemma 2.2.3

The proof directly follows from Lemma 2.2.2 and Lemma 2.2.1. □

Proof of Theorem 2.2.1

For a general form like $y = X\beta + u$, where X has dimension $n \times p$, $\frac{1}{n}X'X$ has full rank and converges to Σ in probability. The solution to ridge regression, i.e., $\arg\min_\beta ||y - X\beta||^2 + v||\beta||_1$, is $\beta_R = (X'X + \nu I_p)^{-1}X'y$. Therefore, $\sqrt{n}(\beta_R -$

$\beta) = -(\frac{1}{n}X'X + \frac{\nu}{n}I_p)^{-1}\frac{\nu}{\sqrt{n}}\beta + (\frac{1}{n}X'X + \frac{\nu}{n}I_p)^{-1}\frac{1}{\sqrt{n}}X'u$. The bias term $-(\frac{1}{n}X'X + \frac{\nu}{n}I_p)^{-1}\frac{\nu}{\sqrt{n}}\beta \rightarrow_p 0$ if $\frac{\nu}{\sqrt{n}} \rightarrow_p 0$. Therefore $\lim_{T\rightarrow\infty}\widetilde{B}_R = B$ holds. □

Proof of Theorem 2.3.1

Let $vec(\hat{R}'_T) = vec(R') + vec(E_R D_T^{-1})$, where E_R is an $m \times m$ matrix, and

$$\begin{aligned}\Psi_T(E_R) &= \left\| vec(\Delta Y) - (Y'_{-1}\widetilde{S} \otimes I_m)vec(R' + E_R D_T^{-1}) \right\|^2_{I_T \otimes \Sigma_u^{-1}} \\ &+ \sum_{i,j=1}^{m} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^\gamma}|R(i,j) + E_R D_T^{-1}(i,j)|\end{aligned}$$

where $\hat{E}_R = \arg\min \Psi_T(E_R)$.

We want to minimize $\Delta_T(E_R) = \Psi_T(E_R) - \Psi_T(0)$.

$$\begin{aligned}\Delta_T(E_R) &= vec(E_R D_T^{-1})'(\widetilde{S}'Y_{-1} \otimes I_m)(I_T \otimes \Sigma_u^{-1})(Y'_{-1}\widetilde{S} \otimes I_m)vec(E_R D_T^{-1}) \\ &- 2vec(U)'(I_T \otimes \Sigma_u^{-1})(Y'_{-1}\widetilde{S} \otimes I_m)vec(E_R D_T^{-1}) \\ &+ \sum_{i,j=1}^{m} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^\gamma}(|R(i,j) + E_R D_T^{-1}(i,j)| - |R(i,j)|) \\ &= vec(E_R)'(D_T^{-1}\widetilde{S}'Y_{-1} \otimes I_m)(I_T \otimes \Sigma_u^{-1})(Y'_{-1}\widetilde{S}D_T^{-1} \otimes I_m)vec(E_R) \\ &- 2vec(\Sigma_u^{-1}UY'_{-1}\widetilde{S}D_T^{-1})'vec(E_R) \\ &+ \sum_{i,j=1}^{m} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^\gamma}(|R(i,j) + E_R D_T^{-1}(i,j)| - |R(i,j)|) \qquad (2.22) \\ &= vec(E_R)'(D_T^{-1}\widetilde{S}'\sum_{t=1}^{T} Y_{t-1}Y'_{t-1}\widetilde{S}D_T^{-1} \otimes \Sigma_u^{-1})vec(E_R) \\ &- 2vec(\sum_{t=1}^{T} \Sigma_u^{-1}u_t Y'_{t-1}\widetilde{S}D_T^{-1})'vec(E_R) \\ &+ \sum_{i,j=1}^{m} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^\gamma}(|R(i,j) + E_R D_T^{-1}(i,j)| - |R(i,j)|)\end{aligned}$$

In Lemma 2.2.1 we see that the first r rows of $\widetilde{S}'$ is a consistent estimator of β'. Thus $\widetilde{R}_1$ is a consistent estimate for α.

Case 1: $0 < r < m$

$$
\begin{aligned}
& \sum_{t=1}^{T} Y_{t-1}Y_{t-1}' = Q^{-1}D_T D_T^{-1} \sum_{t=1}^{T} Z_{t-1}Z_{t-1}' D_T^{-1} D_T Q'^{-1} \\
= \quad & Q^{-1}D_T \left(\begin{array}{cc} T^{-1}\sum_{t=1}^{T} Z_{1,t-1}Z_{1,t-1}' & T^{-3/2}\sum_{t=1}^{T} Z_{1,t-1}Z_{2,t-1}' \\ T^{-3/2}\sum_{t=1}^{T} Z_{2,t-1}Z_{1,t-1}' & T^{-2}\sum_{t=1}^{T} Z_{2,t-1}Z_{2,t-1}' \end{array} \right) D_T Q'^{-1}
\end{aligned}
$$

Let $\widetilde{S} = [\beta + O_p(\frac{1}{T}), \widetilde{S}_2]$ and $Q^{-1} = [q_1, q_2]$. Then, we have

$$
\begin{aligned}
& D_T^{-1}\widetilde{S}' \sum_{t=1}^{T} Y_{t-1}Y_{t-1}' \widetilde{S} D_T^{-1} \\
= \quad & \left[\begin{array}{cc} I_r + O_p(\frac{1}{T}) & \sqrt{T}O_p(\frac{1}{T}) \\ \frac{1}{\sqrt{T}}\widetilde{S}_2' q_1 & \widetilde{S}_2' q_2 \end{array} \right] \left(\begin{array}{cc} T^{-1}\sum_{t=1}^{T} Z_{1,t-1}Z_{1,t-1}' & T^{-3/2}\sum_{t=1}^{T} Z_{1,t-1}Z_{2,t-1}' \\ T^{-3/2}\sum_{t=1}^{T} Z_{2,t-1}Z_{1,t-1}' & T^{-2}\sum_{t=1}^{T} Z_{2,t-1}Z_{2,t-1}' \end{array} \right) \\
& \left[\begin{array}{cc} I_r + O_p(\frac{1}{T}) & \frac{1}{\sqrt{T}} q_1' \widetilde{S}_2 \\ \sqrt{T}O_p(\frac{1}{T}) & q_2' \widetilde{S}_2 \end{array} \right] \qquad (2.23) \\
\rightarrow_d \quad & \left[\begin{array}{cc} \Sigma_{z1z1} & 0 \\ 0 & \widetilde{S}_2' q_2 \Bigg(\Big([0 \quad I_{m-r}] \Sigma_v^{1/2} (\int_0^1 W_m W_m' ds) \Sigma_v^{1/2} \left[\begin{array}{c} 0 \\ I_{m-r} \end{array} \right] \Big)^{-1} \Bigg) q_2' \widetilde{S}_2 \end{array} \right]
\end{aligned}
$$

For the second term in equation (2.22), we have:

$$
\begin{aligned}
& vec(\Sigma_u^{-1}(\sum_{t=1}^{T} u_t Y_{t-1}')\widetilde{S} D_T^{-1}) = vec(\Sigma_u^{-1}(\sum_{t=1}^{T} u_t Y_{t-1}' Q' D_T^{-1}) D_T Q'^{-1} \widetilde{S} D_T^{-1}) \\
= \quad & vec(\left[\begin{array}{cc} T^{-1/2}\sum \Sigma_u^{-1} u_t Z_{1,t-1}' & T^{-1}\sum \Sigma_u^{-1} u_t Z_{2,t-1}' \end{array} \right] \left[\begin{array}{cc} I_r + O_p(\frac{1}{T}) & \frac{1}{\sqrt{T}} q_1' \widetilde{S}_2 \\ \sqrt{T}O_p(\frac{1}{T}) & q_2' \widetilde{S}_2 \end{array} \right]) \\
\rightarrow_d \quad & \left[\begin{array}{c} N(0, \Sigma_{z1z1} \otimes \Sigma_u^{-1}) \\ vec\{\Sigma_u^{-1} Q^{-1} \Sigma_v^{\frac{1}{2}} (\int_0^1 W_m dW_m')' \Sigma_v^{\frac{1}{2}} \left[\begin{array}{c} 0 \\ I_{m-r} \end{array} \right] q_2' \widetilde{S}_2\} \end{array} \right] \qquad (2.24)
\end{aligned}
$$

Next we should pay attention to the last term in eq. (2.22).
For the first r columns of matrix R', the convergence rate of the least square estimator is $\sqrt{T}$. Therefore, if $R(i,j) \neq 0$, $\hat{w}_{i,j} = |\widetilde{R}(i,j)|^{-\gamma} \rightarrow_p |R(i,j)|^{-\gamma}$ and $\sqrt{T}(|R(i,j) + \frac{1}{\sqrt{T}}E_R(i,j)| - |R(i,j)|) \rightarrow sign(R(i,j))|E_R(i,j)|$. By Slutsky's theorem, we have $\frac{\lambda_{i,j,T}^{rank}}{\sqrt{T}} \hat{w}_{i,j} \sqrt{T}(|R(i,j) + \frac{1}{\sqrt{T}}E_R(i,j)| - |R(i,j)|) \rightarrow_p 0$.
If $R(i,j) = 0$, $T^{-\frac{\gamma}{2}}\hat{w}_{i,j} = O_p(1)$ and $\sqrt{T}(|R(i,j) + \frac{1}{\sqrt{T}}E_R(i,j)| - |R(i,j)|) \rightarrow$

$|E_R(i,j)|$. By Slutsky's theorem, we have $\frac{\lambda_{i,j,T}^{rank} T^{\frac{\gamma}{2}}}{\sqrt{T}} T^{-\frac{\gamma}{2}} \hat{w}_{i,j} \sqrt{T}(|R(i,j) + \frac{1}{\sqrt{T}} E_R(i,j)| - |R(i,j)|) \rightarrow_p \infty$.

For the last $m-r$ columns of matrix R', the convergence rate of the least square estimator is T. Therefore, if $T(|R(i,j) + \frac{1}{T} E_R(i,j)| - |R(i,j)|) = |E_R(i,j)|$ and $\frac{\lambda_{i,j,T}^{rank}}{T} T^{\gamma} |T\widetilde{R}(i,j)|^{-\gamma} \rightarrow_p \infty$, where $|T\widetilde{R}(i,j)| = O_p(1)$.

Thus, $\Delta_T(E_R) \rightarrow_d \Delta(E_R)$,where

$$\Delta(E_R) = \begin{cases} vec(E_{R,\mathcal{A}})' M_{\mathcal{A}} vec(E_{R,\mathcal{A}}) - 2W'_{\mathcal{A}} vec(E_{R,\mathcal{A}}) & \text{if } vec(E_R)_k = 0 \quad \forall k \notin \mathcal{A} \\ \infty & \text{otherwise} \end{cases}$$

where $M_{\mathcal{A}} = (\Sigma_{z1z1} \otimes \Sigma_u^{-1})_{\mathcal{A}}$, and $W_{\mathcal{A}} \sim_d N(0, (\Sigma_{z1z1} \otimes \Sigma_u^{-1})_{\mathcal{A}})$. Δ_T is convex and the unique minimum of Δ at $vec(\hat{E}_R)_{\mathcal{A}} = M_{\mathcal{A}}^{-1} W_{\mathcal{A}} \sim_d N(0, (\Sigma_{z1z1} \otimes \Sigma_u^{-1})_{\mathcal{A}}^{-1} (\Sigma_{z1z1} \otimes \Sigma_u^{-1})_{\mathcal{A}} (\Sigma_{z1z1} \otimes \Sigma_u^{-1})_{\mathcal{A}}^{-1})$.

The proof before shows that the non-zero elements in R' can be recognized with this method. However, to prove consistency, we still need to prove that the probability that zero elements can only be selected as non-zero with probability zero, i.e., $\forall k' \notin \mathcal{A}, \lim_{n\to\infty} P(k' \in \mathcal{A}_T^*) = 0$

Suppose $R(i,j) = 0$ but $\hat{R}_T(i,j) \neq 0$, i.e., $k' = jm + i \notin \mathcal{A}$ but $k' \in \mathcal{A}_T^*$. Then according to the Karush-Kuhn-Tucker (KKT for short henceafter) optimality conditions we have

$$X'_{k'}(I_T \otimes \Sigma_u^{-1})(vec(\Delta Y) - Xvec(\hat{R}'_T)) = \frac{1}{2} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^{\gamma}} sign(\hat{R}'_T(i,j)) \tag{2.25}$$

where $X = Y'_{-1}\widetilde{S} \otimes I_m$ and $X_{k'}$ denotes the k' column of X.
Take $T_{k'} = \sqrt{T}$ if $k' \leqslant r$ and $T_{k'} = T$ if $k' > r$. Then divide both sides of the equation above by $T_{k'}$ we get

$$\frac{1}{T_{k'}} X'_{k'}(I_T \otimes \Sigma_u^{-1})(vec(\Delta Y) - Xvec(\hat{R}'_T)) = \frac{1}{T_{k'}} \frac{1}{2} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^{\gamma}} sign(\hat{R}'_T(i,j)) \tag{2.26}$$

If we denote $\widetilde{D}_T = diag[\sqrt{T} I_{mr}, T I_{m(m-r)}]$, then $LHS = \frac{1}{T_{k'}} X'_{k'}(I_T \otimes \Sigma_u^{-1}) vec(U) + \frac{1}{T_{k'}} X'_{k'}(I_T \otimes \Sigma_u^{-1}) X(vec(R') - vec(\hat{R}'_T))$.
From the previous derivation of the asymptotic distribution of $X'(I_T \otimes \Sigma_u^{-1})X$ and $X'(I_T \otimes \Sigma_u^{-1})vec(U)$, we can conclude that LHS is finite in probability.

For the RHS, if $j \leqslant r$, $\frac{\lambda^{rank}_{i,j,T} T^{\frac{1}{2}(\gamma-1)}}{|\sqrt{T}\widetilde{R}(i,j)|^{\gamma}} \to \infty$. If $j > r$, $\frac{\lambda^{rank}_{i,j,T} T^{\gamma-1}}{|T\widetilde{R}(i,j)|^{\gamma}} \to \infty$

By KKT condition, if a zero element is estimated to be nonzero, then the equation (3.46) musts hold. However, the LHS is finite in probability but RHS converges to infinity. Therefore we can exclude this possibility with probability one.

Case 2: $r = 0$
In this case, only the second part of the proof in *Case 1*, i.e. by KKT condition R' can be estimated as non-zero with zero probability.

Case 3: $r = m$
Contrary to *Case 2*, for this case, only the first part of the proof in *Case 1* is necessary.

□

Proof of Theorem 2.3.2

The proof directly follows from Theorem 2.3.1 and Lemma 2.2.3 □

Proof of Theorem 2.3.3

Define $vec(\hat{B}) = vec(B) + vec(\frac{1}{\sqrt{T}}E_B)$ and

$$\begin{aligned}\Psi_T(E_B) &= \left\| vec(\Delta YC) - (C'\Delta X' \otimes I_m)vec(B + \frac{1}{\sqrt{T}}E_B)\right\|^2_{I_T \otimes \Sigma_u^{-1}} \\ &+ \sum_{k=1}^{P}\sum_{i,j=1}^{m} \frac{\lambda^{lag,k}_{i,j,T}}{|\widetilde{B}_{R,k}(i,j)|^{\gamma}} |(B_k(i,j) + \frac{1}{\sqrt{T}}E_{B,k}(i,j)|\end{aligned}$$

where $E_B = [E_{B,1}, \ldots, E_{B,P}]$. Each $E_{B,k}$, $k = 1, \ldots, P$ is an $m \times m$ matrix.
We want to find E_B so as to minimize $\Psi_T(E_B)$. This is equivalent to minimize

$$\begin{aligned}\Psi_T(E_B) - \Psi_T(0) &= vec(\frac{1}{T}E_B)'(\Delta XC\Delta X' \otimes \Sigma_u^{-1})vec(\frac{1}{T}E_B) \\ &- 2vec(\Sigma_u^{-1}UC)'(C'\Delta X' \otimes I_m)vec(\frac{1}{\sqrt{T}}E_B) \\ &+ \sum_{k=1}^{P}\sum_{i,j=1}^{m} \frac{\lambda^{lag,k}_{i,j,T}}{|\widetilde{B}_{R,k}(i,j)|^{\gamma}} \Big(|B_k(i,j) + \frac{1}{\sqrt{T}}E_{B,k}(i,j)| - |B_k(i,j)|\Big)\end{aligned}$$

We have shown the asymptotics of $\frac{1}{T}\Delta XC\Delta X'$ and $\frac{1}{T}UC\Delta X'$ in Lemma 2.A.1. Besides every element in $\widetilde{B}_R$ converges to the true value with rate $\sqrt{T}$, so oracle property argument of adaptive Lasso in Zou (2006) follows.

□

Distribution of $\widetilde{\Pi}$ under Assumption 2.4.1

Lemma 2.A.2. *If error terms u_t in equation* (2.2) *are defined in Assumption 2.4.1, then the least squares estimate for* Π *is distributed as*

$$
\begin{aligned}
& vec\Big[\Big(Q(\widetilde{\Pi}-\Pi)Q^{-1}-[Q\Upsilon\Sigma_{z1}^{-1},0]\Big)D_T\Big] \\
= \quad & vec\Big[[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Qw_tZ'_{1,t-1}, \frac{1}{T}\sum_{t=1}^{T} Qw_tZ'_{2,t-1}]\left[\begin{array}{cc} \frac{1}{T}\sum_{t=1}^{T} Z_{1,t-1}Z'_{1,t-1} & \frac{1}{T^{3/2}}\sum_{t=1}^{T} Z_{1,t-1}Z'_{2,t-1} \\ \frac{1}{T^{3/2}}\sum_{t=1}^{T} Z_{2,t-1}Z'_{1,t-1} & \frac{1}{T^2}\sum_{t=1}^{T} Z_{2,t-1}Z'_{2,t-1} \end{array}\right]^{-1}\Big] \\
\rightarrow_d \quad & \left[\begin{array}{l} N(0,\Sigma_{z1z1}^{-1}\otimes\Sigma_v) \\ vec\Big\{\Big((\Lambda\int_0^1 W_m dW'_m P')' + \sum_{j=1}^{\infty}\Gamma(j)\Big)\left[\begin{array}{c} 0_{r\times(m-r)} \\ I_{m-r}\end{array}\right] \\ \qquad \times\Big(\left[\begin{array}{cc} 0_{(m-r)\times r} & I_{m-r}\end{array}\right]\Lambda(\int_0^1 W_mW'_m ds)\Lambda'\left[\begin{array}{c} 0_{r\times(m-r)} \\ I_{m-r}\end{array}\right]\Big)^{-1}\Big\} \end{array}\right]
\end{aligned}
$$

where W_m is m-dimensional Brownian motion, $D_T = \begin{pmatrix} \sqrt{T}I_r & 0 \\ 0 & TI_{m-r}\end{pmatrix}$, Σ_v is the covariance matrix of $v_t = Qw_t$, $\Lambda = QD(1)P$ with P satisfying $\Sigma_w = PP'$ and $\Gamma(h) = \sum_{j=0}^{\infty} QD_{j+h}\Sigma_w D'_j Q'$.

When the error terms are dependent, the stochastic part $\{u_tZ'_{1,t-1}\}$ is no longer a *martingale difference sequence*. Thus consistency of the least squares estimate does not hold.

To calculate the bias term, we first transform the stationary AR(1) process of $\{Z_{1,t}\}$ into MA(∞) representation. Due to the stationarity of $\{Z_{1,t}\}$, we can derive from

$$\mathcal{G}(L)Z_{1,t} \quad = \quad \beta' u_t, \qquad \text{where } \mathcal{G}(L) = I_r - \beta'\alpha L$$

that

$$Z_{1,t} = \mathcal{G}(L)^{-1}\beta' u_t = \mathcal{G}(L)^{-1}\beta'\kappa(L)w_t \equiv \mathcal{X}(L)w_t$$

Therefore,

$$\frac{1}{T}\sum_{t=1}^{T} Qu_tZ'_{1,t-1} \quad = \quad \frac{1}{T}\sum_{t=1}^{T} Qw_tZ'_{1,t-1} + \frac{1}{T}\sum_{t=1}^{T} Q(\kappa(L)-\kappa(0))w_tZ'_{1,t-1}$$

with $\frac{1}{T}\sum_{t=1}^{T} Q(\kappa(L)-\kappa(0))w_t Z'_{1,t-1} \to_p \sum_{j=1}^{\infty} Q\kappa_j \Sigma_w \mathcal{X}'_{j-1} \equiv Q\Upsilon$. Υ is thus the measure of the correlation between u_t and $Z_{1,t-1}$, which is also the source of bias. Its existence is ensured by the assumption on $\kappa(L)$ and the stationarity of $Z_{1,t}$.
This result leads to a modified version of asymptotic normality as

$$\sqrt{T}vec(\frac{1}{T}\sum_{t=1}^{T} u_t Z'_{1,t-1} - \Upsilon) \to_d N(0, \Sigma_{z1z1} \otimes \Sigma_w)$$

After being corrected for the bias term, the asymptotic distribution has similar form with the *i.i.d* error case. The asymptotics of the unit root process under Assumption 2.4.1 can be referred to Lütkepohl (2007)

□

Proof of Proposition **2.4.1**

The proof is similar to the proof of Theorem 2.3.1 except that the coefficient matrix R is from the QR decomposition of $\Pi + \Upsilon\Sigma_{z1}^{-1}\beta'$, the biased counterpart. The argument with respect to the penalty should be modified as follows.
If at least one element in $R(i,)$ is non-zero, then

$$\begin{aligned}
& \frac{\lambda_{i,T}^{rank}}{||\widetilde{R}(i,)||^{\gamma}}(||R(i,) + \frac{1}{\sqrt{T}}E_R(i,)|| - ||R(i,)||) \\
= & \frac{\lambda_{i,T}^{rank}}{||\widetilde{R}(i,)||^{\gamma}}(||R(i,) + \frac{1}{\sqrt{T}}E_R(i,)|| - ||R(i,)||) \\
= & \frac{\lambda_{i,T}^{rank}}{||\widetilde{R}(i,)||^{\gamma}} \frac{||R(i,) + \frac{1}{\sqrt{T}}E_R(i,)||^2 - ||R(i,)||^2}{||R(i,) + \frac{1}{\sqrt{T}}E_R(i,)|| + ||R(i,)||} \\
= & \frac{\lambda_{i,T}^{rank}/\sqrt{T}}{||\widetilde{R}(i,)||^{\gamma}} \frac{\sum_{j=1}^{m}(2R(i,j) + \frac{1}{\sqrt{T}}E_R(i,j))(E_R(i,j))}{||R(i,) + \frac{1}{\sqrt{T}}E_R(i,)|| + ||R(i,)||} \\
\to_p & \; 0
\end{aligned}$$

If all the elements in $R(i,)$ are zero, then

$$\begin{aligned}
& \frac{\lambda_{i,T}^{rank}}{||\widetilde{R}(i,)||^{\gamma}}(||R(i,) + \frac{1}{T}E_R(i,)|| - ||R(i,)||) \\
= & \frac{\lambda_{i,T}^{rank}T^{\gamma}}{||T\widetilde{R}(i,)||^{\gamma}}||\frac{1}{T}E_R(i,)|| \\
= & \frac{\lambda_{i,T}^{rank}T^{\gamma-1}}{||T\widetilde{R}(i,)||^{\gamma}}||E_R(i,)|| \\
\to & \; \infty
\end{aligned}$$

The left can be finished similar to Wang and Leng (2008). □

Proof of Theorem 2.3.4

As in the proof of Theorem 2.3.1, we define such an objective function:

$$\begin{aligned}
\Psi_T(E) &= \left\| vec(\Delta Y) - (\left[\begin{array}{cc} Y'_{-1}\hat{\beta}^{\dagger} & \Delta X^{p\prime} \end{array}\right] \otimes I_m) vec(\left[\begin{array}{cc} \alpha & B^p \end{array}\right] + \frac{1}{\sqrt{T}}E) \right\|^2_{I_T \otimes \Sigma_u^{-1}} \\
&+ \sum_{i=1}^{m}\sum_{j=1}^{r} \lambda_{i,j,T}^{rank} |\alpha(i,j) + \frac{1}{\sqrt{T}} E_0(i,j)| \qquad (2.27) \\
&+ \sum_{k=1}^{p}\sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_{i,j,T}^{lag,k} |B_k(i,j) + \frac{1}{\sqrt{T}} E_k(i,j)|
\end{aligned}$$

where ΔX^p is the first mp rows of ΔX, $B^p = [B_1, \ldots, B_p]$ and $E = [E_0, E_1, \ldots, E_p]$, E_0 has dimension $m \times r$ and $E_1, \ldots, E_p$ are square matrix of dimension m.

As before, we want to minimize

$$\begin{aligned}
\Delta_T(E) &= \Psi_T(E) - \Psi_T(0) \qquad (2.28) \\
&= vec(\frac{1}{\sqrt{T}}E)'(\left[\begin{array}{c} \hat{\beta}^{\dagger\prime} Y_{-1} \\ \Delta X^p \end{array}\right] \otimes I_m)(I_T \otimes \Sigma_u^{-1})(\left[\begin{array}{cc} Y'_{-1}\hat{\beta}^{\dagger} & \Delta X^{p\prime} \end{array}\right] \otimes I_m) vec(\frac{1}{\sqrt{T}}E) \\
&- 2vec(U)'(I_T \otimes \Sigma_u^{-1})(\left[\begin{array}{cc} Y'_{-1}\hat{\beta} & \Delta X^{p\prime} \end{array}\right] \otimes I_m) vec(\frac{1}{\sqrt{T}}E) \\
&+ \sum_{i=1}^{m}\sum_{j=1}^{r} \lambda_{i,j,T}^{rank} (|\alpha(i,j) + \frac{1}{\sqrt{T}} E_0(i,j)| - |\alpha(i,j)|) \\
&+ \sum_{k=1}^{p}\sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_{i,j,T}^{lag,k} (|B_k(i,j) + \frac{1}{\sqrt{T}} E_k(i,j)| - |B_k(i,j)|)
\end{aligned}$$

Case 1: $0 < r < m$
Because $\hat{\beta}^{\dagger}$ converges to β at the rate of T, we can thus derive the asymptotic distribution of this term:

$$\frac{1}{T}\left[\begin{array}{c} \hat{\beta}^{\dagger} Y_{-1} \\ \Delta X^p \end{array}\right]\left[\begin{array}{cc} Y'_{-1}\hat{\beta}^{\dagger} & \Delta X^{p\prime} \end{array}\right] \rightarrow_p \Sigma_{\Gamma^p\Gamma^p}$$

Based on the proof of Theorem 2.3.1, we can similarly show that

$$\begin{aligned} & (\frac{1}{\sqrt{T}} \begin{bmatrix} \hat{\beta}^{\dagger} Y_{-1} \\ \Delta X^{p} \end{bmatrix} \otimes \Sigma_u^{-1}) vec(U) \\ = \quad & vec(\frac{1}{\sqrt{T}} \Sigma_u^{-1} U \begin{bmatrix} Y'_{-1}\hat{\beta}^{\dagger} & \Delta X^{p'} \end{bmatrix}) \\ \to_d \quad & N(0, \Sigma_{\Gamma^p\Gamma^p} \otimes \Sigma_u^{-1}) \end{aligned}$$

For the penalty imposed on matrix α, $\sum_{i=1}^{m}\sum_{j=1}^{r}\lambda_{i,j,T}^{rank}(|\alpha(i,j) + \frac{1}{\sqrt{T}}E_0(i,j)| - |\alpha(i,j)|) = \sum_{i=1}^{m}\sum_{j=1}^{r}\frac{\lambda_{i,j,T}^{rank}}{\sqrt{T}}(E_0(i,j)sgn(\alpha(i,j))\mathbb{I}(\alpha(i,j) \neq 0) + |E_0(i,j)|\mathbb{I}(\alpha(i,j) = 0))$. By assumption,$\frac{\lambda_{i,j,T}^{rank}}{\sqrt{T}} \to 0$. Therefore, asymptotically, the penalty on α disappears and the estimate is consistent. The same argument works for B_k, $k = 1, \dots, p$.

We have shown that the empirical covariance matrix of the regressors and that between regressor and error terms are all standard as stationary case. The asymptotic distribution in Theorem 2.3.4 follows naturally.

The proof for *Case 2* when $r = 0$ and *Case 3* when $r = m$ are also omitted here.

□

2.B Additional Results

The following lemma recalls the asymptotic distribution of reduced rank regression (see e.g. Lütkepohl (2007), Johansen (1995) and Anderson (2002)).

Lemma 2.B.1. *In special vector error correction model, suppose* $\beta' = [I_r \quad \beta_0']$, *where* β_0' *is of dimension* $(m-r) \times r$. *The estimate from canonical correlation analysis* $\hat{\beta}^{\dagger\prime}$ *has the form* $[\hat{\beta}_1', \hat{\beta}_2']$, *where* $\hat{\beta}_1'$ *are the first* r *columns of* $\hat{\beta}^{\dagger\prime}$.

$$T(\hat{\beta}_2\hat{\beta}_1^{-1} - \beta_0) \to_d (\int_0^1 W_{m-r}^{*} dW_r^{*})'(\int_0^1 W_{m-r}^{*} W_{m-r}^{*'} ds)^{-1} \tag{2.29}$$

where

$$\begin{aligned} W_{m-r}^{*} &= Q^{22}\begin{bmatrix} 0 & I_{m-r} \end{bmatrix} \Sigma_v^{\frac{1}{2}} W_m \\ W_r^{*} &= (\alpha'\Sigma_u^{\frac{1}{2}}\alpha)\alpha'\Sigma_u^{\frac{1}{2}} Q^{-1}\Sigma_v^{\frac{1}{2}} W_m \end{aligned}$$

in which Q^{22} denotes the lower right-hand $(m-r) \times (m-r)$ block of Q^{-1}.

The key point in Lemma 2.B.1 is that W_r^* and W_{m-r}^* are two independent Wiener processes. Thus compared with the term $\Sigma_v^{1/2}(\int_0^1 W_m dW_m')'\Sigma_v^{1/2}\begin{bmatrix} 0_{r\times(m-r)} \\ I_{m-r} \end{bmatrix}$ in Result 1 on page 273 of Lütkepohl (2007), we can see that the distribution in Lemma 2.B.1 is more concentrated around 0. For general VECM, a similar result applies.

2.C Model Specifications for Simulations

Model 2 ($m = 8$, $r = 4$ and $p = 1$)

$$\alpha = \begin{bmatrix} -1.47 & -1.3 & 0 & -1.26 \\ 0 & 0.97 & 0 & 0 \\ 0 & 0 & -0.74 & 0 \\ -1.19 & 0.85 & 0 & 0 \\ -0.55 & 0.78 & -1 & -1.37 \\ 0.8 & 0.75 & 0 & 0 \\ 0 & -0.74 & -1.26 & -0.78 \\ 0 & -1.4 & 0 & 0 \end{bmatrix}$$

$$\beta = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & -0.87 & 1.45 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1.48 \\ 0 & 0 & 1 & 0 & 0 & -1.29 & -0.53 & 0.9 \\ 0 & 0 & 0 & 1 & 0.8 & 1.49 & -0.82 & -0.69 \end{bmatrix}$$

and $B_1 = diag(-0.1852968, 0.4258125, -0.1638084, 0.07833603, -0.5304448, -0.06855371, - 0.7495951, 0.5052671)$.

Model 3 ($m = 8$, $r = 2$, $p = 2$)

$$\alpha = \begin{bmatrix} -0.1608246 & 0.291117 \\ -0.4309348 & -0.2267309 \\ 0.7295761 & 0.7436813 \\ 0.07949743 & -0.5752491 \\ -0.808063 & 0.3370188 \\ -0.9472972 & 0.6852261 \\ -0.8611832 & 0.6208253 \\ 0.8499345 & -0.8429375 \end{bmatrix}$$

$$\beta = \begin{bmatrix} 1 & 0 & 0.1137227 & -0.1445802 & 0.955692 & -0.01119379 & -0.1954843 & -0.9958803 \\ 0 & 1 & -0.4215756 & 0.1502944 & -0.9341822 & -0.5203012 & 0.4701862 & 0.1764804 \end{bmatrix}$$

and $B_1 = diag(0.5013845, 0.1583768, 0.5494133, -0.3385856, 0.2190922, 0.7720483, 0.4980826, 0.02718882)$,
$B_2 = diag(-0.4011076, -0.1267015, -0.4395306, 0.2708685, -0.1752738, -0.6176387, -0.3984661, -0.02175106)$.

Model 4 ($m = 16$, $r = 8$ and $p = 1$)
$B_1 = diag(-0.6148991, 0.168343, 0.3511661, -0.001352618, 0.1055825, 0.05016321, 0.7834411, -0.2399435, -0.1913784, 0.3762232, 0.5340184, 0.4320375, -0.05925948, -0.4302867, 0.6217901, 0.6814101)$ and

$$\Pi = \begin{bmatrix}
-0.2045456 \; 0.127218 \; -0.1044799 \; 0.04996874 \; -0.05324593 \; 0.1565453 \; 0.332533 \; -0.457871 \\
\quad -0.4443822 \; -0.08324072 \; -0.0994021 \; -0.006434139 \; 0.8885221 \; 0.7546155 \; 0.0222507 \; -0.417577 \\
0.02561123 \; -0.2445912 \; -1.076358 \; 0.8504335 \; 0.1481624 \; 0.6820225 \; 0.6595054 \; -1.188968 \\
\quad -0.6543165 \; 0.2423194 \; 0.2819167 \; -0.1265963 \; 1.482206 \; 0.5994158 \; -0.4464372 \; 0.2431477 \\
0.2654349 \; -0.07548686 \; -1.339042 \; 0.2375221 \; -0.2709482 \; 0.2829385 \; 0.4697307 \; -0.7166703 \\
\quad -0.3424121 \; 0.2241369 \; 0.6579697 \; 0.3476774 \; 0.6523763 \; 0.03524423 \; -0.6483029 \; 0.2463741 \\
0.5500683 \; -0.1995099 \; -1.636145 \; -0.05230706 \; 0.8620913 \; 2.380207 \; 0.5911425 \; -0.5798727 \\
\quad -1.777504 \; 0.1451031 \; 1.090046 \; -2.125592 \; 2.355909 \; -0.1184615 \; -0.3810751 \; -0.07006646 \\
0.03690864 \; 0.2959453 \; 0.4596786 \; -0.08504518 \; -0.8577548 \; -0.3276708 \; -0.04811136 \; 0.1974386 \\
\quad 0.1274685 \; 0.3188476 \; -0.158153 \; 0.865952 \; -0.5238296 \; 0.3224605 \; 0.1759896 \; -0.1743132 \\
0.6877773 \; -0.267961 \; -1.200547 \; 0.9718812 \; 0.741968 \; 1.127951 \; 0.3476049 \; -0.6302973 \\
\quad -1.599591 \; 0.08954511 \; 0.6427153 \; -2.008208 \; 1.474142 \; -0.9021317 \; -0.2037194 \; 0.05227726 \\
-0.5995118 \; 0.325451 \; 1.266808 \; -0.6414344 \; -1.09789 \; -1.814652 \; -0.4953283 \; 0.4147672 \\
\quad 2.089613 \; 0.109772 \; -0.6641995 \; 2.750278 \; -2.385913 \; 0.4911569 \; 0.05740444 \; 0.3117873 \\
0.381465 \; -0.04985673 \; -1.095212 \; 0.1829222 \; 0.28933 \; 0.9338472 \; 0.2275248 \; -0.8367844 \\
\quad -0.5197874 \; 0.2886798 \; 0.7498826 \; -0.510993 \; 0.5903355 \; -0.4764813 \; -0.5320649 \; 0.4731749 \\
0.4759285 \; 0.02027912 \; -0.4462453 \; 0.8765776 \; 0.3538885 \; 1.604166 \; 0.3237477 \; -0.9067662 \\
\quad -1.827018 \; 0.3025833 \; 0.1609587 \; -1.733295 \; 1.83846 \; -0.07487888 \; 0.102428 \; -0.09694286 \\
-1.103659 \; 0.3535146 \; 1.854295 \; -1.316152 \; -1.050559 \; -3.093349 \; -0.7909543 \; 1.054735 \\
\quad 2.908839 \; -0.6697658 \; -1.253489 \; 3.332786 \; -3.031778 \; 0.6463785 \; 0.1908991 \; -0.06797553 \\
0.6142871 \; -0.4385424 \; -1.777284 \; 0.4888148 \; 0.8513589 \; 1.79723 \; 0.4217885 \; -0.7512186 \\
\quad -1.715195 \; -0.1673982 \; 0.6688248 \; -2.041544 \; 2.3071 \; -0.5986828 \; -0.5627274 \; 0.3049924 \\
0.4991491 \; -0.3568571 \; -1.473497 \; -0.03773816 \; 1.083164 \; 1.840999 \; 0.4384005 \; -0.1480544 \\
\quad -1.143913 \; 0.1124378 \; 1.153012 \; -1.989919 \; 1.528975 \; -0.4958258 \; -0.3311991 \; 0.06841005 \\
0.3286244 \; 0.1224148 \; 0.2050542 \; -0.06528752 \; -0.2779508 \; -0.1944027 \; -0.4047749 \; 0.200832 \\
\quad 0.4729683 \; 0.3524514 \; 0.2237484 \; 0.347894 \; -1.312519 \; -0.9115838 \; -0.06049354 \; 0.5031275 \\
0.179212 \; -0.06148401 \; -0.2682591 \; 0.002612084 \; 0.2562654 \; 0.6027553 \; 0.06573209 \; 0.06074722 \\
\quad -0.9053709 \; -0.281054 \; -0.04361244 \; -1.034311 \; 1.04103 \; -0.09367657 \; 0.06775278 \; -0.2801906 \\
-0.7085927 \; 0.09905573 \; 1.315568 \; -0.7422261 \; 0.3070841 \; -1.067854 \; -0.4093839 \; 0.7709888 \\
\quad 1.028702 \; -0.6319483 \; -0.7613088 \; 0.3946705 \; -0.9016278 \; 0.4049568 \; 0.4971999 \; -0.4592194 \\
-0.6739596 \; 0.5794677 \; 1.985851 \; -0.7148621 \; -1.103973 \; -1.672337 \; -0.4095454 \; 0.8435712 \\
\quad 1.520876 \; 0.133889 \; -0.8365487 \; 2.135475 \; -2.056529 \; 0.9585998 \; 0.6852929 \; -0.5481826
\end{bmatrix}$$

3 Determination of VECM in High Dimensions

3.1 Introduction

In this chapter, we provide a Lasso-type technique for consistent and numerically efficient model selection when the dimension is allowed to increase with the number of observations at some polynomial rate. Model determination is treated as a joint selection problem of cointegrating rank and VAR lags. In this case, we exploit a sparse model structure in the sense that from a large number of potential cointegration relations, in practice, only a small portion of them are actually prevalent for the system. In the same way, a small and fixed number of VAR lags is considered sufficient for a parsimonious model specification. Within this maximum lag range, however, our model selection technique is independent from the lag ordering detecting non-consecutive lags. We show consistency of model selection by the proposed adaptive group Lasso-VECM estimator requiring only weak moment conditions on the innovations allowing for a wide range of applications. Moreover, we also cover the case of weak dependence in the error term and obtain rank selection consistency despite the fact that least squares pre-estimates of the cointegration matrix are inconsistent in this case. As a by-product, we also derive the statistical properties of the obtained Lasso-estimates for the loadings. A simulation study shows the effectiveness of the proposed techniques in finite samples treating cases of dimension up to 50 with realistic empirical sample sizes. In the empirical example, the new techniques allow us to study a joint system of 15 credit default swaps (CDS) log prices of European sovereign countries and banks - for which there has been no theoretically valid and feasible model determination technique in the literature so far.

Our work builds on the excessive literature of VECM as summarized e.g. in Lütkepohl (2007) as well as on results for Lasso-type techniques from the standard *i.i.d.* setting originating from Tibshirani (1996) and Knight and Fu (2000). In particular, we employ ideas from adaptive Lasso by Zou (2006) for improved selection consistency properties by weighted penalties and use the group structure as in Yuan and Lin (2006) for group-Lasso which allows for simultaneous exclusion and inclusion of certain variables. For the high-dimensional case, consistency results for Lasso have been developed by Bickel et al. (2009) , Zhao and Yu (2006) and in a group-Lasso case in Wei and Huang (2010).

Besides the literature in VECM and Lasso-related methods introduced in the previous chapter, our proposed technique is particularly related to a recent literature which uses Lasso in a high-dimensional time series context. Kock and Callot (2015) and Basu and Michailidis (2015) provide model determination techniques in a stationary high-dimensional VAR context.

There has also been a recent empirical literature which employs Lasso-type penalizing algorithms for VECM without mathematical proofs, see Signoretto and Suykens (2012), Wilms and Croux (2016). To the best of our knowledge, comparable settings of determining cointegrated time series have only been investigated in three recent theoretical papers by Liao and Phillips (2015) in fixed dimensions and Zhang et al. (2018) and Onatski and Wang (2018) in high dimensions. In particular for fixed dimensions, Liao and Phillips (2015) are the first to propose a Lasso-procedure for VECM with theoretical proofs. Their procedure, however, penalizes the eigenvalues of a generally asymmetric matrix, which could introduce complex values but this point is ignored in the theoretical results. Moreover, they fail to propose an algorithm to implement the method in an efficient way. Zhang et al. (2018) provide statistical results for a factor model dealing with high-dimensional non-stationary time series with a focus on forecasting without employing a VECM structure. The focus of Onatski and Wang (2018) is not on model selection consistency but on asymptotic distributional results for testing which require explicit distributional assumptions on the innovations.

The rest of the paper is organized as follows. In Section 3.2 and Section 3.3, we derive the Lasso objective function in a VECM specification in order to determine the cointegration rank and the VAR lags. The consistency results will be derived. Section 3.4 extends the previous econometric analysis to a more general setting with non *i.i.d.* innovations. In Section 3.5 we study the finite-sample performance of the method in several simulation experiments. We also provide an empirical application to CDS data for European countries and banks in Section 3.6. Section 3.7 concludes. All proofs are contained in the Appendix.

Throughout the paper, we use the following notation. For a vector $x \in R^m$, the l_2 norm is defined as $||x||_2 = \sqrt{\sum_{j=1}^m x_j^2}$ and $||x||_\infty = \sup_{1\leqslant j\leqslant m} |x_j|$ is the l_∞ norm. For a matrix $A = ((A_{ij}))$ of dimension $m \times l$, $||A||_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^l A_{ij}^2}$ denotes the Frobenius norm and $||A||_2 = \sup\{||Ax||_2 : x \in R^l \text{ with } ||x||_2 = 1\}$ the l_2 norm. Besides, we denote by $\lambda_j(C)$ the j-th largest eigenvalue of a square matrix C in absolute value, where as $\sigma_j(A)$ is the j-largest singular value of A, i.e. $\sigma_j^2(A) = \lambda_j(A'A)$. Without loss of generality, we assume the sigualr values to be non-negative for notational convenience. We use $vec(A) = [A'_{\cdot 1}, A'_{\cdot 2}, \ldots, A'_{\cdot n}]'$ for vectorizing a matrix A by stacking all columns where $A_{\cdot j}$ is the jth column in matrix A. For

$rank(A) = l < m$, the orthogonal complement to a matrix A is defined as $\mathcal{A}_\perp = \{U \in \mathbb{R}^{m\times(m-l)} | U'A = 0\}$. For an orthonormal $A_\perp$ of A it holds that $A_\perp \in \mathcal{A}_\perp$ and in addition that $A'_\perp A_\perp = I_{m-l}$.

3.2 Cointegration rank selection

3.2.1 Set-up and fundamental results

We consider a general VECM set-up with unknown rank and general lag order which both enter the model selection problem. Thus complete model specification amounts to both rank and lag order determination.

In particular, we consider an m-dimensional $I(1)$ time series Y_t, i.e. Y_t is nonstationary and $\Delta Y_t = Y_t - Y_{t-1}$ is stationary for $t = 1, \ldots, T$ in the following general VECM specification:

$$\Delta Y_t \quad = \quad \Pi Y_{t-1} + B_1 \Delta Y_{t-1} + \cdots + B_P \Delta Y_{t-P} + w_t \tag{3.1}$$

for $t = 1, \ldots, T$, where B_k are $m \times m$ stationary lag coefficient matrices for $k = 1, \ldots, P$ and Π is the $m \times m$ cointegration matrix of rank r with $0 \leqslant r < m$ marking the number of cointegration relations in the system. Π can be decomposed as $\Pi = \alpha\beta'$, where $\beta \in \mathbb{R}^{m\times r}$ constitutes the r long-run cointegrating relations and $\alpha \in \mathbb{R}^{m\times r}$ is a loading matrix of rank r. This decomposition is unique up to a nonsingular matrix H, so only the space of cointegration relations is identified but not β. Without loss of generality, we set β as orthogonal, i.e. $\beta'\beta = I_r$.

Our setup is high-dimensional, thus both, dimension m and cointegration rank r, can grow with sample size T. This treats the practically most important case, as e.g. for large dimensional portfolios with nonstationary components like credit default swaps or exchange rates the number of relevant cointegration relations might increase with sample size. Also from the technical side, this is the interesting innovative case, treating high-dimensionality in the nonstationary parts. For the stationary transient components, however, we set the maximum possible lag length P as sufficiently large but fixed independent of T, such that it is an upper bound for the true lag length p, i.e. $p < P$. In this case, $B_{p+1}, \ldots, B_P$ are all zero matrices. A fixed P or p is chosen for convenience to keep proofs to a minimum with no apparent restriction for practical problems. An extension to P or p increasing with T would be technically straightforward and covered by standard arguments for stationary components (see e.g. Basu and Michailidis (2015)).

In the following, we work with the matrix version of (3.1)

$$\Delta Y = \Pi Y_{-1} + B\Delta X + W \tag{3.2}$$

where $\Delta Y = [\Delta Y_1, \ldots, \Delta Y_T]$, $Y_{-1} = [Y_0, \ldots, Y_{T-1}]$, $B = [B_1, \ldots, B_P]$, $W = [w_1, \ldots, w_T]$, and $\Delta X = [\Delta X_0, \ldots, \Delta X_{T-1}]$ with $\Delta X_{t-1} = \left[\Delta Y'_{t-1}, \ldots, \Delta Y'_{t-P}\right]'$.

For model selection, we disentangle the joint lag-rank selection problem by employing a Frisch-Waugh-idea in the VECM model (3.2). With this, we obtain two independent criteria for lag and rank choice which can be computed separately. For rank selection, the partial least squares pre-estimate $\widetilde{\Pi}$ can be obtained from the corresponding partial model when removing the effect of ΔX in ΔY and Y_{-1} by regressing $\Delta Y M_{\Delta x}$ on $Y_{-1} M_{\Delta x}$ with $M_{\Delta x} = M = I_T - \Delta X'(\Delta X \Delta X')^{-1} \Delta X$. Therefore, (3.2) is equivalent to

$$\Delta \widetilde{Y}_t = \alpha \beta' \widetilde{Y}_{t-1} + \widetilde{w}_t \tag{3.3}$$

with components $\Delta \widetilde{Y} = \Delta Y M$, $\widetilde{Y}_{-1} = Y_{-1} M$ and $\widetilde{W} = WM$. Thus model selection is reduced to rank selection only in (3.3).

Given the high-dimensional set-up, we allow for very general error terms w_t not imposing any specific distributional assumption but just requiring moment assumptions to be satisfied which is key for the practical applicability of the procedure.

Assumption 3.2.1. *For the error component w_t in* (3.1) *exists a representation $w_t = \Sigma_w^{1/2} e_t$ where the elements satisfy the following conditions*

1. *e_t is a sequence of independent copies of e with $E(e) = 0$ and $E(ee') = I_m$ and independence also holds for all elements e_t^k with $k = 1, \ldots, m$.*
2. *Each element in e fulfills $E(|e^k|^{4+\delta}) < \infty$ for some $\delta > 0$ and all $k \leqslant m$.*
3. *For $\Sigma_w = (\Sigma_{w,jk})_{j,k=1}^m$ there exist $\tau_w > 0$ and $0 < K_w < \infty$ such that $\max_{j \leqslant m} \sum_{k=1}^m |\Sigma_{w,jk}| \leqslant K_w$ and $\lambda_m(\Sigma_w) \geqslant \tau_w$.*

The requirement of i.i.d. components in the error term representation allows focusing on the key aspects of our Lasso selection procedure in the high dimensional set-up while keeping technical results to a minimum. In Section 3.4, we show how this Assumption can be generalized admitting linear forms of weak dependence. Such a general setting, however, requires a proof for a general strong invariance principle which is key for our consistency results but not available under weak dependence for high dimensions in the literature so far.

From the first two points in Assumption 3.2.1, Σ_w denotes the covariance matrix of w_t. The third point imposes a sparse structure and ensures positive definiteness of Σ_w through bounding the smallest eigenvalue of Σ_w away from zero. This sparsity condition is satisfied if Σ_w is a banded diagonal matrix with off-diagonal entries far away from the diagonal decaying to zero fast enough (see e.g. Bickel and Levina (2008)). In practice, this seems plausible e.g. in the case of sovereign CDS as treated in the empirical example that geographical distance between countries implies such a cross-section decay structure in the innovations naturally.

Our shrinkage selection procedure for the cointegration rank is based on a least squares pre-estimate of Π from the $M_{\Delta x}$-transformed VECM equation (3.3)

$$\widetilde{\Pi} = \left(\Delta Y M Y_{-1}'\right)\left(Y_{-1} M Y_{-1}'\right)^{-1} \tag{3.4}$$

of the cointegration matrix Π whose statistical properties rely on the decomposition of the transformed $\widetilde{Y}_t$ into a stationary and a non-stationary component. Such a representation generally exists under the following assumptions (see Engle and Granger (1987)):

Assumption 3.2.2. *1. The roots for* $|(1-z)I_m - \Pi z - \sum_{j=1}^{p} B_j (1-z) z^j| = 0$ *is either* $|z| = 1$ *or* $|z| > 1$.

2. The number of roots lying on the unit circle is $m - r$.

3. The matrix $\alpha_\perp'(I_m - \sum_{i=1}^{p} B_i)\beta_\perp$ *is nonsingular with* $||(\alpha_\perp'(I_m - \sum_{i=1}^{p} B_i)\beta_\perp)^{-1}||_2 < \infty$.

The last point of Assumption 3.2.2 is a stronger version than in fixed dimensional case which requires that the smallest singular value of $\alpha_\perp'\beta_\perp$ to be significantly different from zero, which is equivalent to that the basis generating β can not be close to any of the basis of $\alpha_\perp$.

It is well known that for the standard low-dimensional setup with fixed m in (3.1) and Assumptions 3.2.1 and 3.2.2, the standard least squares estimator in (3.4) is consistent (see e.g. Lütkepohl, 2007). In our high-dimensional case, however, we need to explicitly derive its statistical properties. These are key for the construction and validity of a Lasso cointegration rank selection procedure in this paper.

Thus we require the following assumptions reflecting the high-dimensional setting. In the subcase of fixed dimension m, these conditions are trivially fulfilled.

Assumption 3.2.3. *1. All singular values* $\sigma_j(\alpha)$ *of* α *fulfill* $0 < \sigma_r(\alpha) \leqslant \cdots \leqslant \sigma_1(\alpha) < \infty$ *and there exist* $\tau_1 > 0$ *and* $K_1 > 0$ *such that*

$$r^{\tau_1}\sigma_r(\alpha) \geqslant K_1 \ .$$

2. For $B_p = (B_p(i,j))_{i,j=1}^{m}$ *it holds that* $\max_{1 \leqslant i,j \leqslant m} |B_p(i,j)| \geqslant \varepsilon_B > 0$ *with* $\varepsilon_B > 0$ *and for* B *defined in* (3.2) *there exists a positive* $K_B < \infty$ *such that* $||B||_2 < K_B$.

With both dimension m and cointegration rank r increasing with sample size, $\alpha'\alpha$ converges by construction to a compact operator of which the spectrum is well-known to have zero as an accumulation point (cp. Zhao and Yu (2006)). Since therefore the smallest singular value of α in (3.3) has a converging subsequence to zero, Assumption 3.2.3 connects the admissible rate of divergence of the rank r with the rate of decay in singular values of α (cp. the high-dimensional factor model literature, e.g. Li et al. (2017)). Thus for deriving statistical properties of corresponding estimates in this set-up this rate that $\sigma_r(\alpha)$ decays to zero restricts the rate at which r can increase with T. We generally denote elements as relevant if they are non-zero in finite samples but with potentially zero limits or accumulation points asymptotically.

The assumption $||B||_2 < \infty$ is important in a high dimensional setting for avoiding that relevant non-zero elements concentrate on one row or one column only such that a necessary moment bound on ΔY_t can no longer be inferred from the assumptions above.

The statistical properties of $\widetilde{\Pi}$ rely on a Q-transformation of the defining $M_{\Delta x}$-transformed VECM equation (3.3) which allows to disentangle stationary and non-stationary components. We set $Q = \begin{bmatrix} \beta' \\ \alpha_\perp' \end{bmatrix}$ and $Q^{-1} = \begin{bmatrix} \alpha(\beta'\alpha)^{-1} & \beta_\perp(\alpha_\perp'\beta_\perp)^{-1} \end{bmatrix}$, where $\alpha_\perp$ and $\beta_\perp$ are orthogonal complements of α and β respectively, as defined in Assumption 3.2.2. After Q-transformation of (3.3) we get

$$\begin{aligned} \Delta\widetilde{Z}_{1,t} &= \beta'\alpha\widetilde{Z}_{1,t-1} + \widetilde{v}_{1,t} \\ \Delta\widetilde{Z}_{2,t} &= \widetilde{v}_{2,t} \end{aligned} \tag{3.5}$$

where $\widetilde{Z}_t = Q\widetilde{Y}_t = [(\beta'\widetilde{Y}_t)', (\alpha_\perp'\widetilde{Y}_t)']' = \left[\widetilde{Z}_{1,t}' \ \widetilde{Z}_{2,t}'\right]'$ and $\widetilde{v}_t = Q\widetilde{w}_t = [\widetilde{v}_{1t}' \ \widetilde{v}_{2t}']'$. Note that by definition, the first component $\widetilde{Z}_{1,t}$ of dimension r is stationary and the $(m-r)$-dimensional remainder $\widetilde{Z}_{2,t}$ is a unit root process. We also denote $Z_t = QY_t = \left[Z_{1,t}' \ Z_{2,t}'\right]'$, and $v_t = Qw_t = [v_{1t}' \ v_{2t}']'$. From (3.5) the corresponding estimate of the cointegration matrix is obtained as

$$Q\widetilde{\Pi}Q^{-1} = \begin{pmatrix} \sum_{t=1}^T \Delta\widetilde{Z}_{t-1}\widetilde{Z}_{1,t-1}' & \sum_{t=1}^T \Delta\widetilde{Z}_{t-1}\widetilde{Z}_{2,t-1}' \end{pmatrix} \begin{pmatrix} \sum_{t=1}^T \widetilde{Z}_{1,t-1}\widetilde{Z}_{1,t-1}' & \sum_{t=1}^T \widetilde{Z}_{1,t-1}\widetilde{Z}_{2,t-1}' \\ \sum_{t=1}^T \widetilde{Z}_{2,t-1}\widetilde{Z}_{1,t-1}' & \sum_{t=1}^T \widetilde{Z}_{2,t-1}\widetilde{Z}_{2,t-1}' \end{pmatrix}^{-1} \tag{3.6}$$

with $\widetilde{\Pi}$ from (3.4). For this, the statistical properties can be derived in a block-wise way. The result is stated in the following theorem.

Denote by $\mathbf{M} = [\mathbf{M}_1', \mathbf{M}_2']'$ an m-dimensional martingale process with covariance $Q\Sigma_w Q'$ with Σ_w from Assumption 3.2.1 where each component $\mathbf{M}^k$ constitutes a Brownian motion starting at zero and $\mathbf{M}_1$ marks the first subvector of dimension r and $\mathbf{M}_2$ for the vector of the last $m-r$ elements. In the following, given the rank $r < m$, for any matrix $A \in \mathbb{R}^{m\times m}$, denote the top-left $r \times r$ block of A by A_{11}, the bottom-left $(m-r) \times r$ block by A_{12}, the top-right $r \times (m-r)$ block by A_{21}, and the bottom right $(m-r) \times (m-r)$ block by A_{22} respectively.

Theorem 3.2.1. *Let Assumptions 3.2.1, 3.2.2, and 3.2.3 hold. With* $D_T = diag(I_r, TI_{m-r})$

define

$$\widetilde{\Psi} = Q\widetilde{\Pi}Q^{-1}D_T \quad \textit{and} \quad \Psi = \begin{bmatrix} \beta'\alpha & \mathbf{V}_{12} \\ 0 & \mathbf{V}_{22} \end{bmatrix}.$$

with $\mathbf{V}_{i2} = (\int_0^1 d\mathbf{M}_i(s)\mathbf{M}_2'(s))(\int_0^1 \mathbf{M}_2(s)\mathbf{M}_2'(s)ds)^{-1}$ *for* $i = 1, 2$ *with* $\mathbf{M}_1 \in \mathbb{R}^r$ *and* $\mathbf{M}_2 \in \mathbb{R}^{m-r}$ *as defined right above.*

Then for $r = O\left(m^{\frac{1}{2\tau_1+1}}\right)$ *we get blockwise*

$$\begin{aligned} ||\widetilde{\Psi}_{11} - (\beta'\alpha)||_F &= O_p\left(\frac{r}{\sqrt{T}}\right) \\ ||\widetilde{\Psi}_{12} - \mathbf{V}_{12}||_F &= O_p\left(m\sqrt{\frac{(\log T)(\log\log T)^{1/2}}{T^{1/2}}}\right) \\ ||\widetilde{\Psi}_{21}||_F &= O_p\left(\sqrt{\frac{mr}{T}}\right) \\ ||\widetilde{\Psi}_{22} - \mathbf{V}_{22}||_F &= O_p\left(m\sqrt{\frac{(\log T)(\log\log T)^{1/2}}{T^{1/2}}}\right). \end{aligned}$$

Under suitable restrictions on the expansion rates of m and r consistency of all components in $\widetilde{\Psi}$ can be reached. For the stationary components the standard fixed-dimensional $T^{-1/2}$ rate is slowed down by the expansion rates of r and mr. For the nonstationary components, however, the convergence rate depends on the moment conditions of the innovations. In particular, the limit results for the nonstationary blocks in Theorem 3.2.1 yield stochastic elements of Ψ with a general martingale structure of only elementwise Brownian motions instead of a standard multivariate Brownian motion. This is because generally in the high dimensional set-up, a vector composed of elementwise Brownian motion processes does not necessarily follow a multivariate Brownian motion in contrast to standard multivariate fixed dimensional case, see Kosorok and Ma (2007). With higher moment assumptions on the innovation than Assumption 3.2.1, however, a Brownian motion type limit and faster rates of convergences could be achieved. Though for general applicability of our subsequent methodology to financial market data, the stated rates are sufficient and we therefore refrain from imposing moments beyond $4 + \delta$.

Note that the technical condition $m^{\frac{1}{2\tau_1+1}}$ imposes an upper bound for the expansion rate of the rank r depending on the rate of decay of the smallest singular value $\sigma_r(\alpha)$ in T. Combined with Assumption 3.2.3, it implies that for fast decreasing subsequences of $\sigma_r(\alpha)$, the polynomial exponent τ_1 must also be large, imposing a binding restriction on the rate of r. Whereas in the case with any subsequence of $\sigma_r(\alpha)$ approaching zero not too rapidly, identification of relevant elements is easier and thus r can increase faster.

We can combine the blockwise results of Theorem 3.2.1 to obtain the following corollary.

Corollary 3.2.1. *Let Assumptions 3.2.1, 3.2.2, and 3.2.3 hold. Moreover, we require* $m = O(T^{1/4-\varepsilon})$ *with* $\varepsilon \in (0, \frac{1}{4}]$ *and* $r = O\left(m^{\frac{1}{2\tau_1+1}}\right)$*. Then:*

$$||\widetilde{\Psi} - \Psi_0||_F = o_P(1)$$

with $\widetilde{\Psi}$ *as in Theorem 3.2.1 and* $\Psi_0 = Q\Pi Q^{-1} = \begin{pmatrix} \beta'\alpha & 0 \\ 0 & 0 \end{pmatrix} = E(\Psi)$.

Thus the Q-transformed $\widetilde{\Pi}$ consistently estimates the population counterpart under the stated conditions on m and r. The admissible expansion rate $m = O(T^{1/4-\varepsilon})$ mainly results from the mild $(4 + \delta)$ moment condition on the innovations in Assumption 3.2.1 and the strong invariance principle. Fixed dimensions are included as a special case for $\varepsilon = \frac{1}{4}$. Hence, the relevant r-dimensional stationary part can be consistently identified as all other components of Ψ have expectation 0.

3.2.2 Adaptive Group LASSO for rank selection: Idea, procedure and statistical results

The basic principle of standard Lasso-type methods is to determine the number of covariates in a linear model according to a penalized loss-function criterion. Likewise, the determination of the cointegration rank in (3.1) amounts to distinguishing the vectors spanning the r-dimensional cointegration space from the $(m - r)$ basis of its orthogonal complement. This is also equivalent to separating the r relevant singular values of Π in (3.3) from the non-relevant ones, where the number of relevant singular values corresponds to the rank. Thus, the corresponding loading matrix for the stationary part $\widetilde{Z}_{1,t} = \beta'\widetilde{Y}_{t-1}$ in (3.5) is α while the remainder $\beta'_\perp\widetilde{Y}_{t-1}$ should get loading zero in the Q-transformed defining VECM equation (3.3). We use the QR decomposition with column-pivoting[1] to detect the rank of $\Pi = \alpha\beta' = SR$ as the rank of R, where S is orthonormal, i.e. $S'S = I$, and R is an upper triangular matrix [2]. Column-pivoting orders columns in R according to size putting zero rows at the end.[3] Thus the rank r of Π corresponds to the number of relevant columns in R.

[1] We denote the orthogonal matrix in the QR-decomposition by S in order to avoid labeling confusion with the Q-transformation used in equation (3.5)

[2] Such a decomposition exists for any real squared matrix. It is unique for the invertible $\widetilde{\Pi}$ if all diagonal entries of R are fixed to be positive. There are several numerical algorithms like Gram-Schmidt or the Householder reflection which yield the numerical decomposition.

[3] Generally, column pivoting uses a permutation on R such that its final elements $R(i, j)$ fulfill: $|R(1, 1)| \geqslant |R(2, 2)| \geqslant \ldots \geqslant |R(m, m)|$ and $R(k, k)^2 \geqslant \sum_{i=k+1}^{j} R(i, j)^2$. Further properties of this decomposition can be found e.g. in Stewart (1984).

The challenge is, to show that such disentangling of the stationary part $\widetilde{Z}_1$ from the non-stationary $\widetilde{Z}_2$ also works empirically when starting from estimated objects instead of true unobserved population counterparts. Thus calculating the rank from a QR-decomposition with column pivoting of the consistent pre-estimate $\widetilde{\Pi}$ does indeed yield a consistent estimate of the true rank r. In particular, this requires ensuring that true non-relevant singular values, loadings or entries can be distinguished from elements which just appear as non-relevant due to estimation but which in fact truly are relevant which would delude the rank choice. In the following, we show that different speeds of convergence in the stationary and nonstationary parts, however, help to disentangle the two components and can be cleverly exploited in constructing weights for a consistent adaptive group Lasso procedure.

For the Lasso-type objective function, we obtain a pre-estimate for the space of β and $\beta_\perp$ respectively from the QR decomposition with column-pivoting of $\widetilde{\Pi}'$ as

$$\widetilde{\Pi} \quad = \quad \widetilde{R}'\widetilde{S}' = \begin{pmatrix} \widetilde{R}_1' & \widetilde{R}_2' \end{pmatrix} \begin{pmatrix} \widetilde{S}_1' \\ \widetilde{S}_2' \end{pmatrix} = \begin{pmatrix} \widetilde{R}_{11}' & 0 \\ \widetilde{R}_{12}' & \widetilde{R}_{22}' \end{pmatrix} \begin{pmatrix} \widetilde{S}_1' \\ \widetilde{S}_2' \end{pmatrix} \tag{3.7}$$

where $\widetilde{S}$ is $m \times m$ orthonormal, i.e. $\widetilde{S}'\widetilde{S} = I$, with components $\widetilde{S}_1' \in \mathbb{R}^{r\times m}$ and $\widetilde{S}_2' \in \mathbb{R}^{(m-r)\times m}$. $\widetilde{R}$ is an upper triangular matrix with blocks $\widetilde{R}_1 = (\widetilde{R}_{11} \quad \widetilde{R}_{12}) \in \mathbb{R}^{r\times m}$ and $\widetilde{R}_2 = (0 \quad \widetilde{R}_{22}) \in \mathbb{R}^{(m-r)\times m}$ and components with the same notation as for Theorem 3.2.1 where $\widetilde{R}_{11} \in \mathbb{R}^{r\times r}$, $\widetilde{R}_{12} \in \mathbb{R}^{r\times(m-r)}$, and $\widetilde{R}_{22} \in \mathbb{R}^{(m-r)\times(m-r)}$ of $\widetilde{R}$ in (3.7). According to Corollary 3.2.1, for $m = O(T^{1/4-\varepsilon})$ with $\varepsilon \in (0, \frac{1}{4}]$, the estimate $\widetilde{\Pi}$ is a matrix of full-rank and also a consistent estimate of Π. Therefore the lower diagonal elements of $\widetilde{R}_{22}'$ are expected to be small. In particular, they converge to zero asymptotically at unit root speed $1/T$ as is shown in the following Theorem.

Theorem 3.2.2. *Let Assumptions 3.2.1, 3.2.2, and 3.2.3 hold and $\widetilde{R}_1'$ denote the first r and by $\widetilde{R}_2'$ the last $m-r$ columns of $\widetilde{R}'$ in the QR-decomposition* (3.7) *of $\widetilde{\Pi}'$. Besides, define $\tilde{\mu}_k = \sqrt{\sum_{j=k}^m \widetilde{R}(k,j)^2}$. Then for $m = O(T^{1/4-\varepsilon})$ and $r = O(m^{\frac{1}{2\tau_1+1}})$ with $\varepsilon \in (0, \frac{1}{4}]$*

1. $||\beta_\perp'\widetilde{S}_1||_F = O_p(\frac{\sqrt{m}r^{2\tau_1}}{T})$.
2. *$\tilde{\mu}_k$ satisfy*

$$\begin{aligned} \tilde{\mu}_k \quad &\in \quad [\sigma_r(\alpha) - O_p(\sqrt{\frac{mr}{T}}), \sigma_1(\alpha) + O_p(\sqrt{\frac{mr}{T}})] \quad k = 1, 2, \dots, r \\ \tilde{\mu}_k \quad &= \quad O_p(\frac{1}{T}) \quad k = r+1, \dots, m \end{aligned}$$

3. $\max_{1\leqslant j\leqslant r} |\sigma_j(\widetilde{R}_1) - \sigma_j(\alpha)| = O_p(\sqrt{\frac{mr}{T}})$

The first part of Theorem 3.2.2 provides identification of the cointegration space spanned by β. In the respective rate, however, unit root speed is generally slowed down by $\sqrt{m}$ and r^{τ_1} which is larger the faster $\sigma_r(\alpha)$ approaches zero in Assumption 3.2.3. But the subspace distance between $\widetilde{S}_1$ and β converges at a faster rate than the distance between $\widetilde{R}_1$ and α'. This is the key point in order to disentangle stationary and nonstationary components. Moreover, from point 2 of Theorem 3.2.2, the l_2-type weight $\tilde{\mu}_k$ achieves exact unit root speed for the irrelevant parts without affecting identification of the loadings α in speed of convergence. Therefore, $\tilde{\mu}_k$ yields a clearer separation of relevant and irrelevant columns and is the preferred weight for an adaptive Lasso procedure. Note that Theorem 3.2.2 contains the fixed dimensional case as a special case, where identification of the space of β from $\widetilde{S}_1$ is at unit root speed and the standard stationary speed $1/\sqrt{T}$ is obtained for the loadings.

The statistical properties of the QR-components of $\widetilde{\Pi}$ derived in Theorem 3.2.2 inspire the construction of the following adaptive group Lasso objective function (3.8) with group-wise weights for the determination of the cointegration rank (see Wei and Huang (2010) for group Lasso in the standard univariate iid case). Hence columns $\hat{R}'(.,j)$ of the adaptive group-Lasso estimator $\hat{R}'$ minimize the following column-wise criterion over all $R'(.,j)$ for $j = 1, \ldots, m$

$$\sum_{t=1}^{T} \parallel \Delta\widetilde{Y}_t - R'\widetilde{S}'\widetilde{Y}_{t-1} \parallel_2^2 + \sum_{j=1}^{m} \frac{\lambda_T^{rank}}{\tilde{\mu}_j^{\gamma}} ||R'(.,j)||_2 \tag{3.8}$$

where the penalization parameter λ_T^{rank} and the weight γ for adaptiveness in (3.8) are fixed and in practice pre-determined in a data-driven way. See the simulation and application in Sections 3.5 and 3.6 for details. We then obtain an estimate of the true cointegration rank $\hat{r}$ from (3.8) as $\hat{r} = \text{rank}(\hat{R})$, where $\text{rank}(\hat{R})$ equals the number of non-zero columns in $\hat{R}'$.

This adaptive group Lasso procedure (3.8) exploits that according to Theorem 3.2.2 the last $m - r$ columns of $\widetilde{R}'$ converge to zero at a rate faster than the rate of the first r stationary columns for the stated conditions on m and r. With this, we can construct adaptive weights for a model selection consistent group Lasso procedure, which put a faster diverging penalty on any element in the space orthogonal to β and less on those stationary components in the cointegration space.

Remark 3.2.1. *According to Theorem 3.2.2, the subspace distance between $\widetilde{S}_1$ and β converges at a faster rate than the subspace distance of $\widetilde{R}_1$ and α under the given conditions on m and r. Therefore the first step estimation error from using $\widetilde{S}$ in* (3.8) *instead of the infeasible true S_1 is negligible for estimating R from the Lasso criterion.*

Moreover, even when m and r are both fixed, our approach features several advantages compared with existing literature: Firstly, the employed QR-decomposition is always real-valued without further constraints on the matrix $\widetilde{\Pi}$. Thus the Lasso criterion (3.8) only contains real-valued elements and can be minimized with standard optimization techniques. In comparison, a corresponding eigenvalue decomposition of an asymmetric matrix as e.g. in Liao and Phillips (2015) would in general contain complex values leading to a non-standard harmonic function optimization problem in a respective Lasso objective function. Secondly, after the QR-transformation based on the consistent pre-estimator, the objective function (3.8) has the same penalized representation as standard Lasso problem and is therefore straightforward to implement with any available numerically efficient algorithm. So our method is direct and ready to use.

The following theorem provides the statistical properties of adaptive group Lasso estimate from (3.8).

Theorem 3.2.3. *Under Assumptions 3.2.1, 3.2.2, and 3.2.3 and if* λ_T^{rank} *satisfies* $\frac{\lambda_T^{rank}}{\sqrt{T}} r^{\tau_1\gamma+1/2} \to 0$ *and* $\frac{\lambda_T^{rank} T^{\gamma-1}}{m^{3/2}} \to \infty$, $m = O(T^{1/4-\varepsilon})$ *with* $\varepsilon \in (0, 1/4]$, *and* $r = O(m^{\frac{1}{2\tau_1+1}})$. *Then the solution* $\hat{R}$ *of the adaptive group Lasso criterion* (3.8) *satisfies*

1. $\mathbb{P}\Big(\sum_{j=1}^m \mathbb{I}_{\hat{R}'(.,j)\neq 0} = r\Big) \geqslant 1 - C_1 \left(\frac{m^{3/2}}{\lambda_T^{rank} T^{\gamma-1}}\right)^2$ *for some* $C_1 < \infty$.
2. $||\hat{R}_1' - \alpha H||_F = O_p(\sqrt{\frac{mr}{T}})$ *for some orthonormal matrix* H.

Theorem 3.2.3 shows in part 1 rank selection consistency of the adaptive group Lasso technique for all admissible penalties λ_T^{rank} satisfying $\lambda_T^{rank} = o(\frac{\sqrt{T}}{r^{\tau_1\gamma+1/2}})$ and $\frac{m^{3/2}}{T^{\gamma-1}} = o(\lambda_T^{rank})$. Under our assumption on the explosion rate of m and r, setting e.g. γ as 2 allows for a large set of possible λ_T^{rank} choices even if the exact rate of r in unknown. Generally, the best finite sample performance is achieved if γ is not too large as also standard in the literature on stationary adaptive Lasso. Please see also our finite sample results in Section 5.

The lower bound on λ_T^{rank} ensures that with probability approaching 1 the irrelevant groups are excluded by the adaptive group Lasso procedure. Though if λ_T^{rank} increases too rapidly also the non-zero columns of $\hat{R}_1$ will be shrunk to zero. Limiting this bias induces the upper bound on λ_T^{rank}. In total, a larger dimension m decreases the lower bound for the probability that the right model is selected. While a large rank r and small $\sigma_r(\alpha)$ restrict the possible set of λ_T^{rank}, thus impacting the Lasso technique in an indirect way.

In part two of the Theorem, we get as a by-product to consistent cointegration rank selection also consistent estimates for α from the adaptive group Lasso criterion (3.8). Note that the obtained rate of convergence coincides with the infeasible oracle rate in the high-dimensional case when the true cointegration rate was known. In the case of fixed r and m we recover the standard stationary $T^{1/2}$-rate of convergence.

3.3 Lag selection

As for the rank choice, the standard VECM equation (3.2) is transformed in a Frisch-Waugh pre-step in order to focus on the lag selection. In particular, the effect of the nonstationary term Y_{-1} is discarded by employing $C = I_T - Y_{-1}'(Y_{-1}Y_{-1}')^{-1}Y_{-1}$ in (3.2)

$$\Delta\breve{Y}_t = B\Delta\breve{X}_{t-1} + \breve{w}_t \tag{3.9}$$

where we write $\breve{Y} = \Delta YC$ and $\breve{X} = \Delta XC$ with $B = (B_1, \ldots, B_P) \in \mathbb{R}^{m\times mP}$. In contrast to the rank transformation M, the lag transformation C contains nonstationary objects. Thus, the statistical properties of the transformed objects $\breve{Y}_t$ and $\Delta\breve{X}_{t-1}$ must be explicitly derived. For the technical results we refer to Lemma 3.A.3 in the Appendix. For the true lag length $p < P$, we denote by I_B the set of indices with non-zero lag coefficient matrices B_j for $1 \leqslant j \leqslant p$ and set $B_0 \in \mathbb{R}^{m\times lm}$ with $l \leqslant p$ as $B_0 = (B_j)_{j\in I_B}$ the stacked matrix of non-zero lag coefficient matrices in B.

For lag selection, we obtain the least squares estimator $\breve{B}$ and the Ridge estimator $\widetilde{B}$ of the transient lag components B from equation (3.9) as

$$\breve{B} = \left(\frac{1}{T}\sum_{t=1}^{T}\Delta\breve{Y}_t\Delta\breve{X}_{t-1}'\right)\left(\frac{1}{T}\sum_{t=1}^{T}\Delta\breve{X}_{t-1}\Delta\breve{X}_{t-1}'\right)^{-1} \tag{3.10}$$

$$\widetilde{B} = \left(\frac{1}{T}\sum_{t=1}^{T}\Delta\breve{Y}_t\Delta\breve{X}_{t-1}'\right)\left(\frac{1}{T}\sum_{t=1}^{T}\Delta\breve{X}_{t-1}\Delta\breve{X}_{t-1}' + \frac{\lambda_T^{ridge}}{T}I_{mP}\right)^{-1}. \tag{3.11}$$

While we will show that both estimators are consistent, the least squares estimate $\breve{B}$, however, suffers from substantial multicollinearity effects. Therefore it is more favorable in practice to work with the Ridge estimator $\widetilde{B}$. In fact, for the construction of the adaptive Lasso procedure below it is crucial to base the weights on the Ridge pre-estimate $\widetilde{B}$ for valid finite sample selection results on I_B and p.

The statistical properties of $\breve{B}$ and $\widetilde{B}$ are provided in the following Theorem.

Theorem 3.3.1. *Let Assumptions 3.2.1, 3.2.2, and 3.2.3 hold and $\breve{B}$ and $\widetilde{B}$ are as defined in* (3.10) *and* (3.11)*. Assume $m = O(T^{\frac{1}{4}-\varepsilon})$ with $\varepsilon \in (0, 1/4]$. Then*

$$||vec(\breve{B} - B)||_\infty = O_p\left(\sqrt{\frac{\log m}{T}}\right)$$

$$||vec(\widetilde{B} - B)||_\infty = O_p\left(\sqrt{\frac{\log m}{T}}\right)$$

if $\lambda_T^{ridge} = o(\sqrt{T})$ for $\widetilde{B}$ in (3.11).

Note that all components in both estimators depend on the initial transformation C. Therefore for the consistency rates in Theorem 3.3.1 explicit rates of all blocks in (3.10) and (3.11) are crucial and therefore derived in the technical Lemma 3.A.3 in the Appendix.

From Theorem 3.3.1 we obtain consistency results in the l_∞ norm for the vectorized coefficient matrices B_j with $j = 1, \ldots, P$ of the stationary transient components in (3.13). In contrast to the rank selection case, for the stationary lag coefficient pre-estimates there is no difference in speed between true zero coefficient matrices and non-zero ones only estimated as zero as in standard stationary adaptive Lasso selection problems. Thus we adopt the l_∞ norm in order to carefully ensure that if there exists at least one non-zero element in a coefficient matrix, the corresponding lagged term is relevant to the model. Compared to l_2 or Frobenius norm, l_∞ increases with the dimension m only at the logarithmic rate and is independent of the sparsity structure. It is therefore preferred as weight for the adaptive step.

Thus the adaptive Lasso estimate $\hat{B} = (\hat{B}_1, \ldots, \hat{B}_P)$ of the lag coefficient matrices in (3.9) minimizes the following objective function in lag coefficient matrices $B_j \in \mathbb{R}^{m\times m}$ of $B = (B_1, \ldots, B_P) \in \mathbb{R}^{m\times mP}$

$$\sum_{t=1}^{T} ||\Delta\check{Y}_t - \sum_{j=1}^{P} B_j \Delta\check{Y}_{t-j}||_2^2 + \lambda_T^{lag} \sum_{j=1}^{P} ||vec(\widetilde{B}_j)||_\infty^{-\gamma} ||B_j||_F \tag{3.12}$$

As in the case of rank selection, a lag k should be included into the model, whenever $\hat{B}_k$ from the Lasso selection (3.12) is different from zero. Thus, in contrast to other model selection criteria, a Lasso-type procedure allows for the inclusion of non-consecutive lags, which we consider an additional advantage of the procedure.

We obtain an estimate $\hat{p}$ of the true lag length from (3.12) as $\hat{p} = \max_{1\leqslant k\leqslant P}\{k|\hat{B}_k \neq 0\}$. We also define the estimated active set $I_{\hat{B}}$ of (3.12) as the set of indices with non-zero $\hat{B}_j$ for $1 \leqslant j \leqslant p$, i.e., $I_{\hat{B}} = \{j|\hat{B}_j \neq 0\}$ and $\hat{B}_0 = (\hat{B}_j)_{j\in I_B} \in \mathbb{R}^{m\times lm}$ with $l \leqslant p$ consists of estimated coefficient matrices of the true active set I_B.

In the objective function (3.12), we penalize each coefficient matrix jointly by group Lasso rather than penalizing each element in the matrix separately. Such elemen-twise Lasso would be less robust in finite sample performance and potentially lead to problems in economic interpretation.

Remark 3.3.1. *In the adaptive weight, theoretically also the use of the least-squares estimate $\check{B}$ is justified yielding the same consistency result as below for the Ridge estimate $\widetilde{B}$. For a numerically stable adaptive Lasso procedure in finite samples, however, the use of the Ridge weight is essential in order to mitigate the large impact of multicollinearity effects. Also pre-estimates from an elastic net type procedure (see Zou and Hastie (2005)) or sure independence screening (see Fan and Lv (2008)) could be employed for a numerically stable weight in* (3.12). *Their detailed treatment, however, is beyond the scope of this paper.*

The following theorem derives the statistical properties of the adaptive-group Lasso estimates $\hat{B}$ of the lag coefficient matrices.

Theorem 3.3.2. *Let Assumptions 3.2.1, 3.2.2, and 3.2.3 hold. Moreover,* $\frac{\lambda_T^{lag}}{\sqrt{T}} \to 0$ *and* $\frac{\lambda_T^{lag} T^{\frac{1}{2}(\gamma-1)}}{m^2(\log m)^{\gamma/2}} \to \infty$, $m = O(T^{1/4-\varepsilon})$, *then for the solution* $\hat{B}$ *of* (3.12) *with* $I_B, I_{\hat{B}}$ *and* $B_0, \hat{B}_0$ *as defined below* (3.9) *it holds that*

1. $\mathbb{P}\Big(I_{\hat{B}} = I_B\Big) \geqslant 1 - \Big(\frac{m^2(\log m)^{\gamma/2} C_1}{\lambda_T^{lag} T^{1/2(\gamma-1)}}\Big)^2$ *with* $C_1 < \infty$.
2. $||\hat{B}_0 - B_0||_F = O_p(\frac{m}{\sqrt{T}})$.

Theorem 3.3.2 shows lag selection consistency together with consistency of the obtained adaptive Lasso estimates $\hat{B}_0$ for m diverging not too fast. This implies also consistency of the estimated lag length $\hat{p}$ from (3.12). Note that also nonconsecutive lags are identified.

Note that for model selection consistency in the lag there is no impact of the fact that the true rank r is unknown. Technically this is because after C transformation, the effect of the stationary component $Z_{1,t-1}$ is filtered out and the non-stationary $Z_{2,t-1}$ decays to zero. Therefore, the rank just appears in the second order effect, see Lemma 3.A.3 in the Appendix for details.

For consistent lag selection, the tuning parameter must satisfy $\lambda_T^{lag} = o(\sqrt{T})$ and $\frac{m^2(\log m)^{\gamma/2}}{T^{\frac{1}{2}(\gamma-1)}} = o(\lambda_T^{lag})$ with $m = O(T^{1/4-\varepsilon})$. These two conditions correspond to the results from Zou (2006) in the fixed dimensional case. The restrictions on λ_T^{rank} are significantly different from rank selection part for two reasons. First, the denominator of the condition $\frac{m^2(\log m)^{\gamma/2}}{T^{\frac{1}{2}(\gamma-1)}} = o(\lambda_T^{lag})$ is smaller than the corresponding part in the rank selection. This is because the irrelevant basis there converges to zero at the rate of T while in the stationary case, both relevant and irrelevant components converge at the rate of $\sqrt{T}$. This narrows down the possible set of λ_T^{lag} compared to λ_T^{rank}. Second, the largest element in each coefficient matrix must be strictly bounded away from zero so that $\lambda_T^{lag} = o(\sqrt{T})$ is required. Setting γ as 2 or 3, yields good finite sample performance for appropriate choices of λ_T^{lag}. Please see Section 5 for details.

3.4 Rank selection for weakly dependent error terms

In this section, we extend the cointegration rank consistency result to the case of weakly dependent error terms. For our high-dimensional set-up, this requires the derivation of a general functional convergence result under weak dependence which has not been available in the literature so far and is of interest on its own. Moreover, weak dependence also causes pre-estimates for the adaptive Lasso procedure to be biased which is a challenge in the construction of an appropriate rank selection criterion.

To derive and illustrate the main points, we focus in this section on the simple VECM case only with no lags (See also e.g. Phillips (2014) in the fixed dimensional case). Thus we work with

$$\Delta Y_t \quad = \quad \Pi Y_{t-1} + u_t \tag{3.13}$$

for $t = 1, \dots, T$ where the dimension m of Y_t and rank r of $\Pi = \alpha\beta'$ are diverging with T as in (3.1). But now, we allow for a general weakly dependent form of the error term u_t in (3.13).

Assumption 3.4.1. *The error term has the representation* $u_t = \sum_{j=0}^{\infty} A_j w_{t-j}$ *with* $A_0 = I_m$ *where for the components it holds that*

1. w_t *satisfies Assumption 3.2.1.*
2. *the coefficient matrices satisfy* $\sum_{j=1}^{\infty} j||A_j||_F < \infty$.

In Assumption 3.4.1, the coefficient matrices of this infinite moving average process u_t must decay to zero fast enough so that u_t is a weakly dependent multiple time series and thus the partial sums can still be approximated by a Wiener process element-wise. In particular, we get the following functional convergence result.

Theorem 3.4.1. *Let Assumption 3.4.1 hold. Then each element in* u_t *has bounded* $(4+\delta)$*-th moment as the original innovation* e_t*. Besides, the partial sum of each* u_t^k *can be approximated by Brownian motion, i.e.,*

$$\max_{s \leqslant T} |\sum_{t=1}^{s} u_t^k - \mathbf{M}^k(s)| = O_{a.s.}(T^{1/4}(\log T)^{3/4}(\log\log T)^{1/2}), \quad k = 1, 2, \dots, m$$

where each component $\mathbf{M}(s)^k$ *in* $\mathbf{M}(s)$ *follows a Brownian motion starting at zero and the covariance matrix of* $\mathbf{M}(1)$ *is* $\Sigma_u = (I_m + \sum_{j=1}^{\infty} A_j)\Sigma_w(I_m + \sum_{j=1}^{\infty} A_j)'$.

This theorem is the crucial element for deriving the statistical properties of the adaptive Lasso pre-estimates and consistency of the cointegration rank selection procedure.

We can directly obtain the least-squares estimator $\widetilde{\Pi}$ of Π for the simple VECM (3.13) as

$$\widetilde{\Pi} = (\sum_{t=1}^{T} \Delta Y_t Y_{t-1}')(\sum_{t=1}^{T} Y_{t-1} Y_{t-1}')^{-1} \tag{3.14}$$

which coincides with the estimate from equation (3.4) for the no lag case $p = 0$.

We derive its statistical properties by using the Q-transformation from (3.5) to distinguish the r stationary Z_1 from the $m-r$ nonstationary components Z_2 in $Z = QY = (Z_1', Z_2')'$. Note that the Q-transformed problem (3.13) simplifies in the rank only case to

$$\begin{array}{rcl} \Delta Z_{1,t} & = & \beta'\alpha Z_{1,t-1} + v_{1,t} \\ \Delta Z_{2,t} & = & v_{2,t} \end{array} \tag{3.15}$$

with $v_t = Qu_t = (v_{1,t}', v_{2,t}')'$ where $v_1 \in \mathbb{R}^r$ and $v_2 \in \mathbb{R}^{m-r}$. Note that here $\mathbf{E}(v_t Z_{1,t-1}')$ is non-zero, due to the possible dependence in u_t and thus in v_t according to Assumption 3.4.1. This causes an endogenity bias such that left subspace generated by $\widetilde{\Pi}$ in (3.14) does no longer approximate α but $\alpha_\star$ defined as

$$\begin{array}{rcl} \alpha_\star' & = & \alpha' + \Sigma_{z1}^{-1}\Gamma_{v1z1}^{1\prime}(\alpha'\beta)^{-1}\alpha' + \Sigma_{z1}^{-1}\Gamma_{v2z1}^{1\prime}(\beta_\perp'\alpha_\perp)^{-1}\beta_\perp' \\ & = & \alpha' + \Sigma_{z1}^{-1}\Gamma_{uz1}^{1\prime} \end{array} \tag{3.16}$$

with $\Gamma_{uz1}^1 = \mathbf{E}(u_t Z_{1,t-1}')$ and $\Sigma_{z1} = \mathbf{E}(Z_{1,t-1}Z_{1,t-1}')$. We also set $\Gamma_{viz1}^1 = \mathbf{E}(v_{it}Z_{1,t-1}')$ with $i \in \{1,2\}$. Though, for $\alpha_\star$ defined in (3.16) Assumption 3.2.3 is not sufficient to ensure non-singularity of $\alpha_\star'\alpha_\star$. Singularity, however, would affect rank selection consistency of the Lasso procedure since the estimation error for the relevant r basis would be inflated by an exactly zero smallest singular value of $\alpha_\star$. We therefore require the condition in part 1 of Assumption 3.2.3 not only for α but also for the biased object $\alpha_\star$. This is needed even in fixed dimensional case where an $\alpha_\star$ without full row-rank would increase the estimation error for $\widetilde{S}_1$ in the QR-decompsiton (3.7) from unit root speed $\frac{1}{T}$ in Theorem 3.2.2 to only $\frac{1}{\sqrt{T}}$ which makes it indistinguishable from the stationary parts. Therefore we require the following assumption

Assumption 3.4.2. *Let part 1 of Assumption 3.2.3 hold. Moreover, the singular values of $\alpha_\star$ satisfy $0 < \sigma_r(\alpha_\star) \leqslant \cdots \leqslant \sigma_1(\alpha_\star) < \infty$. And there exist $K_2 > 0$ and $\tau_2 > 0$ such that $r^{\tau_2}\sigma_r(\alpha_\star) \geqslant K_2$.*

The size of τ_2 and τ_1 restricts the admissible expansion rates in r and m as shown in the Theorems below. For the rest of the subsection, we assume wlog that $\tau_2 \geqslant \tau_1$. The other cases would be easier to be identified.

Let $\mathbf{M}(s)$ denote the m-dimensional martingale process defined in Theorem 3.4.1, where $\mathbf{M}_1(s)$ marks the first r elements and $\mathbf{M}_2(s)$ the last $m-r$ components.

Theorem 3.4.2. *Let Assumptions 3.2.2, 3.4.1 and 3.4.2 hold. With $D_T = diag\,(I_r,\ TI_{m-r})$ and $\widetilde{\Pi}$ from (3.14) define*

$$\widetilde{\Psi} = Q\widetilde{\Pi}Q^{-1}D_T.$$

Moreover, denote

$$\Psi_\star = \left[\begin{array}{cc} \beta'\alpha + \Gamma_{v1z1}^1\Sigma_{z1}^{-1} & \Gamma_{v1z1}^1\Sigma_{z1}^{-1}\Xi(\int_0^1 \mathbf{M}_2(s)\mathbf{M}_2'(s)ds)^{-1} + \mathbf{V}_{12} \\ \Gamma_{v2z1}^1\Sigma_{z1}^{-1} & \Gamma_{v2z1}^1\Sigma_{z1}^{-1}\Xi(\int_0^1 \mathbf{M}_2(s)\mathbf{M}_2'(s)ds)^{-1} + \mathbf{V}_{22} \end{array}\right]$$

where $\Xi = (\beta'\alpha)^{-1}\Big((\beta'\alpha + I_r)\Gamma^{1,\prime}_{v2z1} + \Sigma_{v1v2} + \Gamma^0_{12} + \int_0^1 d\mathbf{M}_1(s)\mathbf{M}_2'(s)\Big)$ *and* $\mathbf{V}_{ij} = (\int_0^1 d\mathbf{M}_i(s)\mathbf{M}_j'(s) + \Gamma^0_{ij})(\int_0^1 \mathbf{M}_j(s)\mathbf{M}_j'(s)ds)^{-1}$ *for* $i,j = 1,2$ *with* $\Gamma^0 = \sum_{k=1}^{\infty}\mathbf{E}(v_t v_{t-k}')$ *and all other elements as defined below* (3.16).
Then for $r = O(m^{\frac{1}{2\tau_1+1}})$ *it holds that*

$$\begin{aligned}
||\widetilde{\Psi}_{11} - \Psi_{\star,11}||_F &= O_p(\frac{r}{\sqrt{T}}) \\
||\widetilde{\Psi}_{12} - \Psi_{\star,12}||_F &= O_p(m\sqrt{\frac{(\log T)^{3/2}(\log\log T)}{T^{1/2}}}) \\
||\widetilde{\Psi}_{21} - \Psi_{\star,21}||_F &= O_p(\sqrt{\frac{mr}{T}}) \\
||\widetilde{\Psi}_{22} - \Psi_{\star,22}||_F &= O_p(m\sqrt{\frac{(\log T)^{3/2}(\log\log T)}{T^{1/2}}})\ .
\end{aligned}$$

There are two main differences between this result and the independent case in Theorem 3.2.1. First, there is a bias term $\Gamma^1_{vz1} \neq 0$ due to the correlation between u_t and Z_{t-1}. Second, the rate of convergence for the unit root part is slightly smaller due to the larger exponent in the $\log T$-term. Though, the driving denominator is still $T^{1/4}$ as before. Moreover, the rate restriction on r coincides with the iid case since the inverse of $\beta'\alpha$ in Ξ causes the l_2-norm of Ξ and thus of $\Psi_{\star,12}, \Psi_{\star,22}$ to increase at rate of r^{τ_1}.

For the parts in the QR-representation of $\widetilde{\Pi}$ we find the following key separation into stationary and nonstationary components

Theorem 3.4.3. *Let Assumptions 3.2.2, 3.4.1 and 3.4.2 hold and* $\widetilde{R}_1'$ *denote the first* r *and by* $\widetilde{R}_2'$ *the last* $m-r$ *columns of* $\widetilde{R}'$ *in the QR-decomposition* (3.7) *of* $\widetilde{\Pi}'$ *in* (3.14). *With* $\tilde{\mu}_k = \sqrt{\sum_{j=k}^m \widetilde{R}(k,j)^2}$ *for* $m = O(T^{1/4-\varepsilon})$ *and* $r = O(m^{\frac{1}{2\tau_2+1}})$ *where* $\varepsilon \in (0, \frac{1}{4}]$ *it holds that*

1. $||\beta_\perp'\widetilde{S}_1||_F = O_p(\frac{\sqrt{m}r^{\tau_1+2\tau_2}}{T})$.
2. $\tilde{\mu}_k$ *satisfy*

$$\begin{aligned}
\tilde{\mu}_k &\in [\sigma_r(\alpha_\star) - O_p(\sqrt{\frac{mr}{T}}), \sigma_1(\alpha_\star) + O_p(\sqrt{\frac{mr}{T}})] \quad k = 1,2,\dots,r \\
\tilde{\mu}_k &= O_p(\frac{r^{\tau_1}}{T}) \quad k = r+1,\dots,m
\end{aligned}$$

3. $\max_{1\leqslant j\leqslant r}|\sigma_j(\widetilde{R}_1) - \sigma_j(\alpha_\star)| = O_p(\sqrt{\frac{mr}{T}})$.

Theorem 3.4.3 shows that identification of the cointegration space occurs at a slightly slower speed of convergence as in the iid-case of Theorem 3.2.2. Weak dependence in the innovation also slows down the convergence of the Lasso adaptive weights in the true zero parts from unit root speed to $\frac{r^{\tau_1}}{T}$. Both points make it harder for adaptive Lasso (3.8) to disentangle true stationary and nonstationary components. Technically, the difference in convergence rates of Theorem 3.4.3 and Theorem 3.2.2 results from the fact that for $\Psi_\star$ with the additional bias Γ^1_{viz1} the l_2 bounds for blocks in Ψ cannot be attained. Convergence in the third part can only be attained for $\alpha_\star$ instead of α but the rate is unaffected.

Therefore, the same logic for the design of group Lasso weights from the iid case can still be employed. Thus, we can still use the adaptive group Lasso objective function (3.8) for rank selection with a pre-estimate $\widetilde{S}$ from a QR-decomposition of $\widetilde{\Pi}$ in (3.14). As before, it yields a columnwise estimate of $\hat{R}'$ from which we can determine the cointegration rank. The statistical properties of this procedure are provided in the following theorem.

Theorem 3.4.4. *Under Assumptions 3.2.2, 3.4.1 and 3.4.2, if* λ_T^{rank} *satisfies* $\frac{\lambda_T^{rank}}{\sqrt{T}} r^{\tau_2\gamma+1/2} \to 0$ *and* $\frac{\lambda_T^{rank} T^{\gamma-1}}{m^{3/2} r^{\tau_1(\gamma+1)}} \to \infty$, $m = O(T^{1/4-\varepsilon})$ *with* $\varepsilon \in (0, 1/4]$, *and* $r = O(m^{\frac{1}{2\tau_2+1}})$, *then the solution* $\hat{R}$ *of the adaptive group Lasso criterion* (3.8) *with pre-estimate* $\widetilde{S}$ *from a QR-decomposition of* $\widetilde{\Pi}$ *in* (3.14) *satisfies*

1. $\mathbb{P}\Big(\sum_{j=1}^m \mathbb{I}_{\hat{R}'(,j)\neq 0} = r\Big) \geqslant 1 - \bar{C}_0(\frac{m^{3/2} r^{\tau_1(\gamma+1)}}{\lambda_T^{rank} T^{\gamma-1}})^2$ *for some* $\bar{C}_0 < \infty$

2. $||\hat{R}_1' - \alpha_\star H||_F = O_p(\sqrt{\frac{mr}{T}})$

for some orthonormal matrix H.

Theorem 3.4.4 shows that given our assumptions, even if the innovations are weakly dependent, rank selection is still consistent. The estimate of the loading matrix, however, only consistently identifies $\alpha_\star$ as defined in (3.16) which generally differs from α.

3.5 Simulations

In this section, we illustrate the finite sample performance of our adaptive Lasso methodology. We consider three different high-dimensional scenarios

1. dimension $m = 20$, rank $r = 5$ and lag $p = 1$
2. dimension $m = 20$, rank $r = 5$ and lag $p = 0$
3. dimension $m = 50$, rank $r = 10$ and lag $p = 0$

Exact model specifications of Π in (3.1) are constructed randomly by first generating two orthonormal matrices $U, V \in \mathbb{R}^{m\times r}$. Such orthonormal matrices can be obtained from QR-decomposition or singular value decomposition of a matrix with each element drawn from a standard normal distribution. Then we randomly draw elements for an $r \times r$ diagonal matrix Λ from univariate standard normal until $\Pi = U\Lambda V'$ first satisfies Assumption 3.2.2. As the main focus of this paper is rank selection in a cointegrated model, in all set-ups coefficient matrices B_j are set as diagonal with elements also drawn from a univariate standard normal. In this section, we set $P = 3$ to reduce computational time. Innovations w_t in (3.1) are drawn from the standard Normal or t-distribution with degrees of freedom $df \in \{8, 20, 200\}$ fulfilling the moment condition of Assumption 3.2.1. We study different degrees of cross-sectional dependence, with banded covariance matrices of the innovations of the form $\Sigma_w = (\rho^{|i-j|})_{ij}$ for $\rho = 0.0, 0.2, 0.4, 0.6$. We consider different combinations of these parameters for sample sizes $T = 400, 800, 1200, 1600$.

The exact specification of the considered setting and the estimating procedure can be replicated from the R-code available at `https://github.com/liang-econ/High_Dimensional_Cointegration` by setting the same seed. Throughout this section, the tuning parameter $\lambda_T^{rank}(\lambda_T^{lag})$ is selected by BIC as follows

$$\min_{\lambda} \log |\hat{\Sigma}_w(\lambda)| + \frac{\log T}{T} ||vec(A(\lambda))||_0 \tag{3.17}$$

where $A = \hat{R}(\lambda)$ in rank selection and $A = \hat{B}(\lambda)$ in lag selection, and $\hat{\Sigma}_w(\lambda)$ denotes the sample covariance matrix of the residuals for λ from (3.3) or (3.9).

In the following tables, each cell contains the percentages XX/YY of correct model selections by solving (3.8) and (3.12) for $b = 100$ repetitions of the respective model, where XX denotes the number of correct rank selections while YY is the number of correct lag length identifications. When the model has no transient terms, there exists only one number XX representing rank selection results.

Table 3.1 studies the performance of the adaptive group Lasso procedure for $m = 20$ dimensions with true rank $r = 5$ and lag $p = 1$ with $\rho = 0$ in the cross-correlation of the innovations. From top to bottom the difficulty of the selection problem increases with less existing moments in the innovation terms. This is also reflected in the reported results with excellent overall performance except in extreme cases where $T^{1/4}$ is smaller than 5, but the treated dimension is $m = 20$. Here, the conditions for Lasso selection consistency with $m = o(T^{1/4})$ are hard to justify. Though performance of the Lasso procedure is still quite good but affected by heavier tails in the innovations in particular in the lag selection case. For the same setup of Π and B as in Table 3.1, we report model selection results for an almost normal type of innovation with $df = 200$ and substantial tail thickness $df = 20$ across different levels of strength in the cross-sectional correlation Σ_w in Table 3.2. The results show that even for substantial correlation with $\rho = 0.6$, performance is reliable for $T \geqslant 800$ even in the case of for $df = 20$ innovations with excess-kurtosis of 0.375. Generally, a larger degree of freedom leads to better rank selection results given

	$T = 400$	$T = 800$	$T = 1200$	$T = 1600$
$N(0, I_m)$	84/98	100/100	100/100	100/100
$df = 200$	81/96	100/100	100/100	100/100
$df = 20$	76/98	100/100	100/100	100/100
$df = 8$	82/99	100/100	100/100	100/100

Table 3.1. Model selection results for model 1 with $m = 20$, rank $r = 5$, lag $p = 1$, $\rho = 0$ and $\gamma = 3$.

	$T = 400$	$T = 800$	$T = 1200$	$T = 1600$
$df = 200, \rho = 0.0$	81/96	100/100	100/100	100/100
$df = 200, \rho = 0.2$	78/98	100/100	100/100	100/100
$df = 200, \rho = 0.4$	80/97	100/100	100/100	100/100
$df = 200, \rho = 0.6$	71/88	97/100	100/100	100/100
$df = 20, \rho = 0.0$	76/98	100/100	100/100	100/100
$df = 20, \rho = 0.2$	91/97	100/100	100/100	100/100
$df = 20, \rho = 0.4$	85/96	100/100	100/100	100/100
$df = 20, \rho = 0.6$	59/80	96/100	100/100	100/100

Table 3.2. Model selection results for model 1 with $m = 20$, rank $r = 5$, lag $p = 1$ and $\gamma = 3$

the same T and ρ. Besides, simulations show that the size of ρ has a significant effect on model selection, which highlights the importance of Assumption 3.2.1 on the structure of Σ_w, i.e, the column-wise sums of absolute values must converge fast enough.

Note that Tables 3.1 and 3.2 are obtained for $\gamma = 3$. Table 3.3 shows the effect of γ on model selection in finite sample in the same setting of model 1. In small samples, $\gamma = 3$ is generally the best choice for consistent rank and lag selection. But with $\gamma = 2$ only slighly weaker results are obtained, while larger choices increase the weight in the penalty too much and yield substantially less appealing results across all considered tail specifications, cross-correlations and samples sizes. Generally, in the case of model 1 with 20 dimensions and $r = 5$, $p = 1$, the results demonstrate that with a sample size of $T = 800$ we get 100% perfect rank selection across all cross-correlation and tail scenarios given non-Gaussian innovations. Compare this to usual simulation evidence in high-dimensional set-ups as e.g in Zhang et al. (2018) which exclusively use Gaussian innovations and require sample sizes of $T = 2000$ for comparable performance.

Besides, we present the estimation error of the loading matrix $\hat{R}_1$ and the cointegrating space $\tilde{S}_1$ in Figure 3.1 for $df = 20$ and in Figure 3.2 for $df = 200$ in the case $\rho = 0.0$. Because α and β are only unique up to rotation, the estimation error here is measured by using orthogonal projection matrices to uniquely identify subspace

		$df = 20$		$df = 200$	
		$T = 400$	$T = 800$	$T = 400$	$T = 800$
$\gamma = 2$	$\rho = 0.0$	89/91	98/100	89/92	99/100
	$\rho = 0.4$	78/82	97/100	75/88	94/100
$\gamma = 3$	$\rho = 0.0$	76/98	100/100	81/96	100/100
	$\rho = 0.4$	85/96	100/100	80/97	100/100
$\gamma = 4$	$\rho = 0.0$	46/99	100/100	48/97	100/100
	$\rho = 0.4$	50/99	100/100	48/97	100/100

Table 3.3. Model selection results for model 1 with $m = 20$, rank $r = 5$ and lag $p = 1$ for different ρ cross-section dependence with different γ-choices.

distances. In particular, we employ the R package LDRTools based on average orthogonal projection matrices proposed by Liski et al. (2016). The left bar in each plot corresponds to $T = 800$ and the right one to $T = 1200$. The estimation error for the cointegrating space is significantly smaller than that for the loading matrix due to the faster rate of convergence. Moving from sample size 800 to 1200 significantly improves results in both cases.

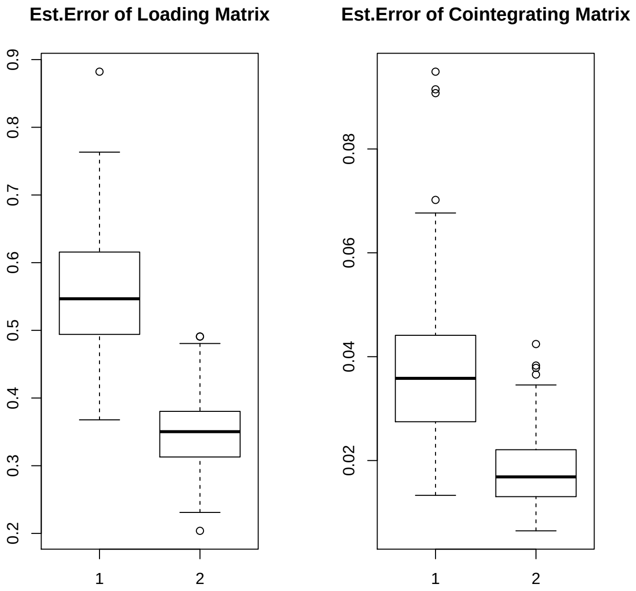


Figure 3.1. Estimation Error of model 1 ($m = 20$, $r = 5$, $p = 1$) with t-distributed innovations and $df = 20$ for $\rho = 0$ setting $\gamma = 3$. Results are shown for $T = 800$ marked as case 1 on the x-axis and for case 2 of $T = 1200$

Model 2 uses the same Π as model 1 but considers only rank selection in VECM

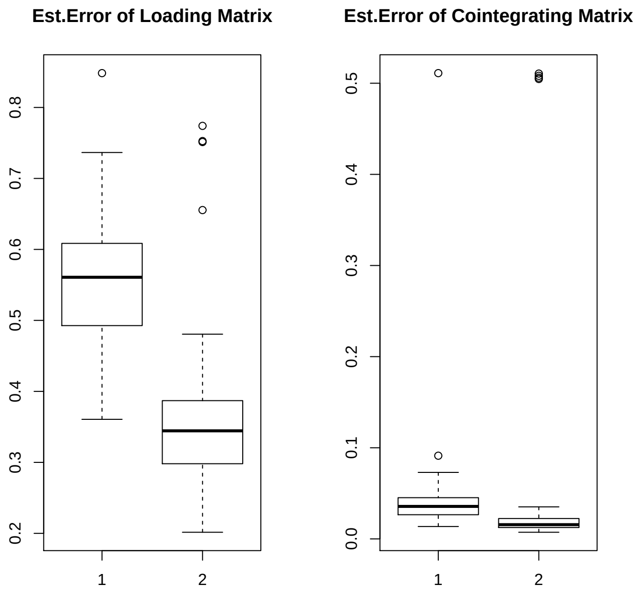


Figure 3.2. Estimation Error of model 1 ($m = 20$, $r = 5$, $p = 1$) with t-distributed innovations and $df = 200$ for $\rho = 0$ setting $\gamma = 3$. Results are shown for $T = 800$ marked as case 1 on the x-axis and for case 2 of $T = 1200$

without transient dynamics, i.e. setting $B = 0$. Thus the problem is simpler and technically, the step of the Frisch-Waugh transformation by M in (3.3) can be omitted. The results can be found in Table 3.4. In small samples with $T = 400$ and for large ρ, this provides improvements in comparison to 3.2. Thus without lags, we get satisfactory performance even in these challenging cases of strong cross-sectional dependence.

To test the performance of our method in case of weakly dependent innovations, we generate the weakly dependent innovations according to a MA(2) process. The innovations in the underlying MA process are *i.i.d.* generated from t-distribution with degree of freedom 20 and 200 respectively. The weakly dependent innovations are generated by

$$u_t = w_t + A_1 w_{t-1} + A_2 w_{t-2}$$

where w_t follows t-distribution with covariance $\Sigma_w = (\rho^{|i-j|})_{ij}$ as defined before. Besides, $A_1 = (a_{1,ij}) = (0.8^i \mathbb{I}_{i=j})$ and $A_2 = (a_{2,ij}) = ((-0.4)^i \mathbb{I}_{i=j})$ satisfy Assumption 3.4.1. As in Table 3.1, we set $\gamma = 3$ and choose λ by BIC. See Table 3.5 for results. When $T \geqslant 800$, the rank selection results are satisfactory, which is consistent with the theoretical results.

	$T = 400$	$T = 800$	$T = 1200$	$T = 1600$
$df = 200, \rho = 0.0$	100	100	100	100
$df = 200, \rho = 0.2$	98	100	100	100
$df = 200, \rho = 0.4$	96	100	100	100
$df = 200, \rho = 0.6$	75	100	100	100
$df = 20, \rho = 0.0$	98	100	100	100
$df = 20, \rho = 0.2$	97	100	100	100
$df = 20, \rho = 0.4$	94	100	100	100
$df = 20, \rho = 0.6$	74	100	100	100

Table 3.4. Model selection result for model 2 with $m = 20$, rank $r = 5$, lag $p = 0$ and $\gamma = 3$

	$T = 400$	$T = 800$	$T = 1200$	$T = 1600$
$df = 200, \rho = 0.0$	100	100	100	100
$df = 200, \rho = 0.4$	86	99	100	100
$df = 20, \rho = 0.0$	99	100	100	100
$df = 20, \rho = 0.4$	94	100	100	100

Table 3.5. Rank selection result for model 2 with $m = 20$, rank $r = 5$, lag $p = 0$ and $\gamma = 3$ and weakly dependent innovations.

In Table 3.6, we present the rank selection results for the 50-dimensional case of model 3. Compare this to the usual simulation scenarios the high-dimensional non-stationary time series literature which usually do not go beyond dimension 20 (see e.g. Zhang et al. (2018)). We focus on results for innovations following a t-distribution with $df = 20$ and $df = 200$ respectively, with $\rho = 0.0$, i.e. $\Sigma_w = I_m$ only. For both cases, when $T \geqslant 2000$, the true rank can be estimated almost 100% correct. The increased sample size reflects the difficulty of the problem in dimensionality.

For the high-dimensional set-ups treated before, there exists no other valid feasible method for model determination against which we could evaluate our technique. Therefore, although our techniques are tailored to the high-dimensional case, we briefly illustrate that they can also be employed in standard low dimensions where benchmarks exist. In particular, we compare our methods with the Lasso-type

	$T = 800$	$T = 1200$	$T = 1600$	$T = 2000$	$T = 2400$
$m = 50, df = 20$	51	64	89	93	99
$m = 50, df = 200$	55	78	95	97	100

Table 3.6. Rank selection result for $m = 50$ with t-distributed innovations. $\rho = 0$ and $\gamma = 3$.

techniques in Liao and Phillips (2015) using the "hardest" of their 2-dimensional models treated with $r = 1$ and $p = 3$. In particular, we set $\Pi = \begin{pmatrix} -1 & -0.5 \\ 1 & 0.5 \end{pmatrix}$ and $B_1 = B_3 = \text{diag}(0.4\ 0.4)$, $B_2 = 0$ and $\Sigma_w = \text{diag}(1.25\ 0.75)$. With 5000 simulation replications we get the following model selection results: for $T = 100$ we get 100%/86.14% while for $T = 400$ we obtain 100%/99.96% which compare to 99.54%/99.80% and 100%/99.98% by Table 2 in Liao and Phillips (2015). In their other settings, we also found similar comparable performance of the two techniques. Results are omitted here for the sake of brevity but are available on request.

3.6 Empirical Example

In this section, we employ our method to study the interconnectedness of the European sovereign and key players of the banking system during and after the financial crisis. We use CDS log prices of ten European countries and five selected financial institutions provided by Bloomberg terminal: *Germany, France, Belgium, Austria, Denmark, Ireland, Italy, Netherland, Spain, Portugal, BNP Paribas, SocGen Bank, LCL Bank, Danske Bank, Santander Bank* [4]. The sovereign countries we choose have different debt levels. The considered time span is from $Jan.1, 2013$ to $Dec.31, 2016$ with 1041 observations. *BNP Paribas, SocGen Banks* are chosen because they rank among the top three Europe based investment banks in Euro-Zone revenues. The other three banks are selected across EU countries covering the whole span from north to south and representing the variety of different financial market and general economic conditions. Initial Augumented Dicky Fuller tests show that the 15 variables are non-stationary but the first-order differences are stationary.

Figure 3.4 suggests that there exits a strong co-movement among these components. Using our Lasso procedure, we find that there exist two cointegration relations. Figure 3.3 gives an impression on the stable time evolution of these cointegrated series. Moreover, the time when the cointegrated series exhibit extreme values coincides with some important economic events. For example, in the middle of the year 2013, European countries were bargaining over the solution for the sovereign debt crisis while at the beginning of 2016 there occurred an economic slowdown in the key global economies.

To present the inter-connections among these 15-dimensional VECM components, we calculate the forecast error variance decomposition (FEVD henceafter) due to the cointegrated part, i.e., the forecast error variance decomposition[5] derived from (3.3). From Table 3.7 reporting the FEVC results for a 5-and 10-step forecast horizon, we

[4] UK is excluded due to Brexit

[5] see e.g. Section 2.3.3 of Lütkepohl (2007)

can conclude that leading economies in European Union, such as Germany, are neither risk-exporter nor risk-importer in the whole system.) Italy is the largest risk-exporter among the considered sovereign countries and Spain ranks second. Moreover, Italy and Spain have significant mutual influence on each other. The banks have stronger interconnectedness among themselves than with the sovereign countries. Moreover, Figure 3.5 shows the contribution of Italy to the FEVD of other variables in the full horizon from step 0 to 30, which is consistent with the results in Table 3.7.

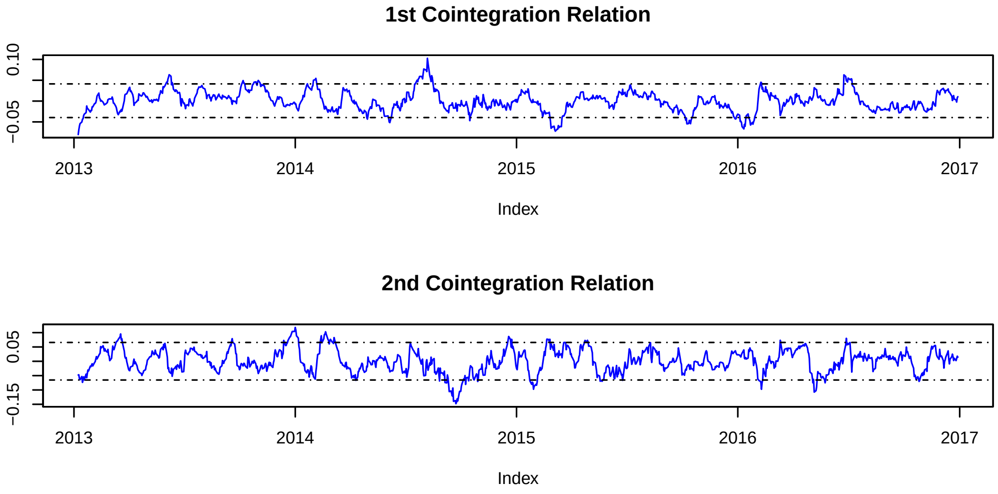


Figure 3.3. Significant cointegration relations

	DE	FR	BE	AT	DK	IE	IT	NL	ES	PT	BNP	SOCGEN	LCL	DAN	SANTAN
DE_5	96.03	2.27	0.02	0.28	0.20	0.01	0.21	0.07	0.62	0.02	0.03	0.14	0.05	0.02	0.04
FR_5	0.84	95.83	0.16	1.25	0.01	0.08	0.26	0.07	0.45	0.10	0.20	0.30	0.09	0.12	0.24
BE_5	0.00	0.26	97.95	0.08	0.09	0.54	0.02	0.55	0.01	0.08	0.08	0.24	0.05	0.06	0.00
AT_5	0.34	1.77	0.07	95.34	0.31	0.45	0.02	0.66	0.68	0.08	0.13	0.00	0.02	0.02	0.13
DK_5	0.17	0.01	0.11	0.37	97.79	0.05	0.00	1.12	0.02	0.04	0.16	0.05	0.03	0.01	0.07
IE_5	0.04	0.04	0.29	0.23	0.02	98.07	0.66	0.35	0.03	0.09	0.03	0.01	0.08	0.05	0.01
IT_5	0.06	0.29	0.00	0.00	0.00	0.60	82.39	0.00	13.22	0.40	0.26	1.10	0.41	0.10	1.15
NL_5	0.09	0.16	0.45	0.51	0.77	0.64	0.01	96.74	0.01	0.01	0.01	0.00	0.01	0.24	0.36
ES_5	0.44	0.72	0.05	0.12	0.03	0.32	14.83	0.00	82.17	1.01	0.04	0.04	0.02	0.02	0.21
PT_5	0.01	0.00	0.16	0.02	0.05	0.74	1.53	0.03	1.28	94.58	0.06	0.10	0.08	0.31	1.06
BNP_5	0.03	0.02	0.07	0.00	0.03	0.13	0.02	0.00	0.02	0.00	87.81	6.59	3.62	0.10	1.54
SOCGEN_5	0.07	0.07	0.02	0.02	0.02	0.06	0.28	0.01	0.01	0.08	6.19	87.63	4.22	0.15	1.19
LCL_5	0.01	0.05	0.05	0.10	0.00	0.35	0.49	0.01	0.40	0.01	10.93	13.18	70.87	0.23	3.33
DAN_5	0.01	0.05	0.04	0.00	0.00	0.06	0.17	0.10	0.00	0.09	0.04	0.10	0.15	99.05	0.14
SANTAN_5	0.00	0.17	0.00	0.02	0.02	0.01	0.68	0.20	0.15	0.25	3.36	2.60	1.90	0.28	90.35
Sum_5	2.09	5.87	1.48	3.01	1.54	4.03	19.18	3.16	16.91	2.26	21.52	24.45	10.73	1.69	9.47
DE_10	96.00	2.28	0.02	0.28	0.20	0.01	0.22	0.07	0.63	0.03	0.03	0.14	0.04	0.02	0.04
FR_10	0.80	95.92	0.15	1.25	0.01	0.07	0.24	0.07	0.40	0.11	0.20	0.31	0.09	0.12	0.25
BE_10	0.00	0.26	97.96	0.08	0.09	0.54	0.02	0.55	0.01	0.07	0.08	0.24	0.05	0.06	0.00
AT_10	0.35	1.78	0.06	95.27	0.31	0.44	0.02	0.66	0.71	0.09	0.14	0.00	0.02	0.02	0.14
DK_10	0.17	0.01	0.11	0.37	97.79	0.05	0.00	1.12	0.02	0.04	0.16	0.04	0.03	0.01	0.07
IE_10	0.04	0.04	0.29	0.23	0.02	98.15	0.62	0.35	0.02	0.07	0.03	0.00	0.09	0.05	0.01
IT_10	0.06	0.30	0.00	0.00	0.00	0.58	82.18	0.00	13.32	0.40	0.28	1.15	0.44	0.10	1.19
NL_10	0.10	0.16	0.45	0.51	0.77	0.65	0.01	96.72	0.00	0.00	0.01	0.00	0.00	0.24	0.36
ES_10	0.46	0.73	0.06	0.11	0.03	0.33	14.97	0.00	81.93	1.07	0.03	0.04	0.01	0.02	0.20
PT_10	0.01	0.00	0.16	0.02	0.05	0.74	1.54	0.03	1.29	94.57	0.06	0.10	0.07	0.31	1.06
BNP_10	0.03	0.02	0.08	0.00	0.03	0.13	0.02	0.00	0.02	0.00	88.27	6.39	3.44	0.10	1.47
SOCGEN_10	0.07	0.07	0.01	0.02	0.02	0.07	0.28	0.01	0.01	0.08	6.07	87.93	4.07	0.15	1.16
LCL_10	0.01	0.05	0.05	0.11	0.00	0.38	0.53	0.01	0.43	0.00	11.44	13.80	69.44	0.23	3.52
DAN_10	0.01	0.05	0.04	0.00	0.00	0.07	0.18	0.10	0.00	0.09	0.03	0.09	0.13	99.09	0.13
SANTAN_10	0.00	0.17	0.00	0.02	0.02	0.01	0.66	0.20	0.14	0.25	3.33	2.57	1.87	0.28	90.48
Sum_10	2.10	5.92	1.49	3.01	1.55	4.08	19.29	3.16	17.01	2.30	21.87	24.87	10.37	1.70	9.57

Table 3.7. Each cell implies the contribution of variable denoted by its column name to the forecast error variance of the variable denoted by its row name. The number in row names is the horizon of the FEVD. The row denoted by *Sum* calculates the sum of each column except the element on the diagonal, which is the total contribution to all the other variables.

3.7 Conclusion

This paper discusses how to determine high dimensional VECM under quite general assumptions. It proposes a general groupwise adaptive Lasso procedure which is easily implementable and thus ready to use for practitioners. We show that it works under quite general assumptions such as mild moment conditions on the innovations while rank and dimension can increase with sample size T. In particular, consistency results in rank and lag selection are obtained for dimension m satisfying $m = O(T^{1/4-\varepsilon})$ for some small and positive ε. Besides, we also derive the statistical properties of the estimator in case of weakly dependent innovations. According to our best knowledge, this paper is the first to provide a theoretically justified solution to model determination of VECM in a high-dimensional set-up. Questions like efficient estimation of the cointegrating space and faster diverging rates in the dimension require different approaches and thorough investigation. They are therefore left for future research.

3.A Proofs

Technical Lemmas

Lemma 3.A.1. *Let $A \in \mathbb{R}^{m\times k}, B \in \mathbb{R}^{k\times n}$ with $\sigma_1(A), \sigma_1(B)$ denoting the largest singular value. $C \in \mathbb{R}^{m\times m}$ is non-singular with largest/smallest singular value denoted by $\sigma_1(C)/\sigma_m(C)$ Given T observations the estimators for A, B, C are denoted as $\widetilde{A}, \widetilde{B}, \widetilde{C}$ and satisfy*

$$||\widetilde{A} - A||_F = O(q(T)), \qquad ||\widetilde{B} - B||_F = O(q(T)), \qquad ||\widetilde{C} - C||_F = O(q(T))$$

with $q(T) \to 0$ as $T \to \infty$, then

$$\begin{aligned} ||\widetilde{A}\widetilde{B} - AB||_F &= O(\max(\sigma_1(A), \sigma_1(B))q(T)) \\ ||\widetilde{C}^{-1} - C^{-1}||_F &= O(\sigma_m^{-2}(C)q(T)) \end{aligned}$$

Proof. By the Mirsky version of matrix perturbation theory (see Theorem 4.11 of Stewart and Sun (1990)), the singular values of the estimated matrix are consistent for those of the true matrix, i.e.,

$$|\sigma_j(\widetilde{A}) - \sigma_j(A)| = O(q(T))$$

Therefore,

$$\begin{aligned} ||\widetilde{A}\widetilde{B} - AB||_F &= ||(\widetilde{A} - A)\widetilde{B} + A(\widetilde{B} - B)||_F \\ &\leqslant ||(\widetilde{A} - A)||_F||\widetilde{B}||_2 + ||A||_2||(\widetilde{B} - B)||_F \\ &= O(\sigma_1(B)q(T) + \sigma_1(A)q(T)) \end{aligned}$$

The argument can be proved by showing that

$$\begin{aligned} ||\widetilde{C}^{-1} - C^{-1}||_F &= ||\widetilde{C}^{-1}(C - \widetilde{C})C^{-1}||_F \\ &\leqslant ||\widetilde{C}^{-1}||_2||(\widetilde{C} - C)||_F||C^{-1}||_2 \\ &= O(\sigma_m(C)^{-2}q(T)) \end{aligned}$$

□

Lemma 3.A.2. *Under Assumptions 3.2.1, 3.2.2 and 3.2.3, and $m = (T^{1/4-\varepsilon})$ with $\varepsilon \in (0, 1/4]$ the following results hold:*

1. $||\frac{1}{T}\sum_{t=1}^T \Delta Y_t \Delta X'_{t-1}||_2 = O_P(1)$ *and* $||\frac{1}{T}\sum_{t=1}^T Y_{t-1}\Delta X'_{t-1}||_2 = O_P(1)$.
2. $||\frac{1}{T}\sum_{t=1}^T \Delta X_{t-1}\Delta X'_{t-1}||_2 = O_P(1)$ *and* $||(\frac{1}{T}\sum_{t=1}^T \Delta X_{t-1}\Delta X'_{t-1})^{-1}||_2 = O_P(1)$.
3. $||\frac{1}{\sqrt{T}}\sum_{t=1}^T w_t \Delta X'_{t-1}||_2 = O_P(1)$ *and thus* $||\frac{1}{T}\sum_{t=1}^T \widetilde{w}_t\widetilde{w}'_t - \Sigma_w||_F = O_p(\frac{m}{\sqrt{T}})$.

Proof. To simplify the analysis, we rewrite the general VAR process in (3.2) as a VAR(1) process by defining:

$$\begin{aligned} F_t^1 &= [Y_t', \Delta Y_t', \ldots, \Delta Y_{t-p+1}']' \\ F_t^0 &= [Z_{1t}', \Delta Y_t', \ldots, \Delta Y_{t-p+1}']' \end{aligned}$$

Then we get from (3.2) and the stationary components after Q-transformation of (3.2) that

$$F_t^1 = \begin{pmatrix} \Pi + I_m & B_1 & \ldots & B_{p-1} & B_p \\ \Pi & B_1 & \ldots & B_{p-1} & B_p \\ 0 & I_m & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & I_m & 0 \end{pmatrix} F_{t-1}^1 + \begin{pmatrix} w_t \\ w_t \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.18}$$

$$F_t^0 = \begin{pmatrix} \beta'\alpha + I_r & \beta' B_1 & \ldots & \beta' B_{p-1} & \beta' B_p \\ \alpha & B_1 & \ldots & B_{p-1} & B_p \\ 0 & I_m & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & I_m & 0 \end{pmatrix} F_{t-1}^0 + \begin{pmatrix} \beta' w_t \\ w_t \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.19}$$

Setting the matrix in (3.18) as Φ_1, the cointegrated process F_t^1 has the compact VAR(1) representation

$$F_t^1 = \Phi_1 F_{t-1}^1 + [w_t', w_t', 0_{m(p-1)}']' = \sum_{j=0}^{\infty} \Phi_1^j [w_{t-j}', w_{t-j}', 0_{m(p-1)}']'.$$

The VMA(∞) representation holds as the $m(p+1)$-dimensional square matrix Φ_1 has $m-r$ eigenvalues on the unit circle and all the others within the unit circle due to Assumptions 3.2.2 and 3.2.3. In a similar way, denoting by Φ_0 the matrix in (3.19) we get

$$F_t^0 = \Phi_0 F_{t-1}^0 + [v_{1t}', w_t', 0_{m(p-1)}']' = \sum_{j=0}^{\infty} \Phi_0^j [v_{1,t-j}', w_{t-j}', 0_{m(p-1)}']'$$

with $\lambda_1(\Phi_0) < 1$ and $||\Phi_0||_2 < \infty$. Define $\tilde{v}_t = [v_{1,t-j}', w_{t-j}', 0_{m(p-1)}']'$ with covariance matrix $\Sigma_{\tilde{v}}$, then with $m = (T^{1/4-\varepsilon})$ for some finite K large enough:

$$||\frac{1}{T}\sum_{t=1}^{T} F_t^0 F_t^{0\prime} - \mathbf{E}(F_t^0 F_t^{0\prime})||_2 + O_P(m/\sqrt{T})$$

while

$$\begin{aligned}
||\mathbf{E}(F_t^0 F_t^{0\prime})||_2 &= ||\sum_{j=0}^{\infty} \Phi_0^j \Sigma_{\tilde{v}} \Phi_0^{j\prime}||_2 \\
&\leqslant \sum_{j=0}^{\infty} ||\Phi_0^j||_2^2 ||\Sigma_{\tilde{v}}||_2 \\
&\leqslant ||\Sigma_{\tilde{v}}||_2 \left(\sum_{j=0}^{K} ||\Phi_0^j||_2^2 + \sum_{j=K+1}^{\infty} ||\Phi_0^j||_2^2 \right)
\end{aligned}$$

where $||\Sigma_{\tilde{v}}||_2$ is bounded due to $\lambda_1(\Sigma_w) < \infty$ by Assumption 3.2.1. Moreover, $\sum_{j=0}^{K} ||\Phi_0^j||_2^2$ is bounded for finite K and $\sum_{j=K+1}^{\infty} ||\Phi_0^j||_2^2$ is bounded due to Gelfand's formula, since $||\Phi_0^K||^{1/K} \leqslant \lambda_1(\Phi_0) + \epsilon(K) < 1$ for sufficiently large K. Thus $||\frac{1}{T}\sum_{t=1}^{T} F_t^0 F_t^{0\prime}||_2 = O_P(1)$ which implies points 1. and 2. in the Lemma.

For the l_2- norm in part 3, it remains to show that $||\frac{1}{T}\sum_{t=1}^{T} Z_t \Delta X_{t-1}^{\prime}||_2 = O_P(1)$. Note that the empirical covariance matrix between $Z_{1,t-1}$ and ΔX_{t-1} is part of the covariance matrix of F_t^0 and thus bounded. Therefore the claim follows if $||\frac{1}{T}\sum_{t=1}^{T} Z_{2,t-1} \Delta X_{t-1}^{\prime}||_2 = O_P(1)$. After Q-transformation of (3.2), we obtain for all non-stationary components

$$Z_{2,t} = Z_{2,t-1} + \alpha_{\perp}^{\prime} B \Delta X_{t-1} + v_{2,t}$$

Define a stationary series $f_t = [v_{2,t}^{\prime}, \Delta X_{t-1}^{\prime}]^{\prime}$ and its partial sum $F_{t-1}^2 = \sum_{s=1}^{t-1} f_s$. Then

$$\begin{aligned}
F_t^2 F_t^{2\prime} &= (F_{t-1}^2 + f_t)(F_{t-1}^2 + f_t)^{\prime} \\
&= F_{t-1}^2 F_{t-1}^{2\prime} + F_{t-1}^2 f_t^{\prime} + f_t F_{t-1}^{2\prime} + f_t f_t^{\prime}
\end{aligned} \tag{3.20}$$

By summing up (3.20) and dividing both sides by T, we get

$$\frac{1}{T}\sum_{t=1}^{T} F_{t-1}^2 f_t^{\prime} + \frac{1}{T}\sum_{t=1}^{T} f_t F_{t-1}^{2\prime} = (\frac{1}{\sqrt{T}} F_T^2)(\frac{1}{\sqrt{T}} F_T^{2\prime}) - \frac{1}{T}\sum_{t=1}^{T} f_t f_t^{\prime}$$

has bounded l_2 norm due to the stationarity of f_t. Therefore $\frac{1}{T}\sum_{t=1}^{T} Z_{2,t-1}\Delta X_{t-1}^{\prime}$ has bounded l_2 norm since $Z_{2,t-1} = [I_{m-r}, \alpha_{\perp}^{\prime} B] F_{t-1}^2$.

Moreover, it holds that

$$\begin{aligned}
& ||\frac{1}{T}\sum_{t=1}^{T}\widetilde{w}_t\widetilde{w}_t' - \Sigma_{ww}||_F \\
= \; & ||\frac{1}{T}\sum_{t=1}^{T}w_tw_t' - \Sigma_{ww} - (\frac{1}{T}\sum_{t=1}^{T}w_t\Delta X_{t-1}')(\frac{1}{T}\Delta X_{t-1}\Delta X_{t-1}')^{-1}(\frac{1}{T}\sum_{t=1}^{T}\Delta X_{t-1}w_t')||_F \\
\leqslant \; & ||\frac{1}{T}\sum_{t=1}^{T}w_tw_t' - \Sigma_{ww}||_F + \frac{1}{T}||(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}w_t\Delta X_{t-1}')(\frac{1}{T}\Delta X_{t-1}\Delta X_{t-1}')^{-1}(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Delta X_{t-1}w_t')||_F \\
= \; & O_p(\frac{m}{\sqrt{T}}) + O_p(\frac{m}{T}) = O_p(\frac{m}{\sqrt{T}})
\end{aligned}$$

as the first term in the second to last line $||\frac{1}{T}\sum_{t=1}^{T}w_tw_t' - \Sigma_{ww}||_F = O_P(\frac{m}{\sqrt{T}})$ due to a standard law of large numbers for stationary time series. For the last term in that line, note that the l_2-norm of the expression inside the norm is $O_P(1)$, which implies that the stated Frobenius-norm is at most $O_P(m)$. □

Proof of Theorem 3.2.1

Proof. For the claims of the theorem, it is sufficient to show that

$$\begin{aligned}
& \left\|\frac{1}{T}\sum_{t=1}^{T}\Delta\widetilde{Z}_t\widetilde{Z}_{t-1}' - \begin{bmatrix} (\beta'\alpha)\Sigma_{z1.\Delta X} & -\Sigma_{v1v2} \\ 0 & \int_0^1 d\mathbf{M}_2(s)\mathbf{M}_2(s)' \end{bmatrix}\right\|_F . \\
= \; & O_p\Big(\sqrt{\frac{r^2}{T}} + \sqrt{\frac{mr}{T}} + \sqrt{\frac{m^2(\log T)(\log\log T)^{1/2}}{T^{1/2}}}\Big) \qquad (3.21)
\end{aligned}$$

and

$$\begin{aligned}
& \left\|D_T^{-1}\frac{1}{T}\sum_{t=1}^{T}\widetilde{Z}_{t-1}\widetilde{Z}_{t-1}' - \begin{bmatrix} \Sigma_{z1.\Delta X} & -(\beta'\alpha)^{-1}\big(\Sigma_{v1v2} + \int_0^1 d\mathbf{M}_1(s)\mathbf{M}_2'(s)\big) \\ 0 & \int_0^1 \mathbf{M}_2(s)\mathbf{M}_2'(s)ds \end{bmatrix}\right\|_F \\
= \; & O_p\Big(\sqrt{\frac{r^2}{T}} + \sqrt{\frac{mr}{T}} + \sqrt{\frac{m^2(\log T)(\log\log T)^{1/2}}{T^{1/2}}}\Big) \qquad (3.22)
\end{aligned}$$

where $\Sigma_{z1.\Delta x} = \Sigma_{z1} - \Sigma_{z1\Delta x}\Sigma_{\Delta x}^{-1}\Sigma_{\Delta xz1}$ and $\Sigma_{v1v2} = \beta'\Sigma_w\alpha_\perp$.

We show (3.22) and (3.21) by studying blockwise elements of $\frac{1}{T}\sum_{t=1}^{T}\Delta\widetilde{Z}_t\widetilde{Z}_{t-1}'$ and $D_T^{-1}\frac{1}{T}\sum_{t=1}^{T}\widetilde{Z}_{t-1}\widetilde{Z}_{t-1}'$. Thus according to (3.6) we need to consider the following 8 different blocks.

1.+2. purely stationary blocks $b_{11} = \frac{1}{T}\Delta Z_1 M Z'_{1,-1}$ *and* $\chi_{11} = \frac{1}{T} Z_{1,-1} M Z'_{1,-1}$
For the second block the standard law of large numbers argument from Lemma 3.A.2 yields:

$$\frac{1}{T}\sum_{t=1}^{T}\widetilde{Z}_{1,t-1}\widetilde{Z}'_{1,t-1} = \Sigma_{z1} - \Sigma_{z1\Delta x}\Sigma^{-1}_{\Delta x}\Sigma_{\Delta x z1} + R_1 \tag{3.23}$$

with $||R_1||_F = O_P(r/\sqrt{T})$. For the first term we get from (3.5) we get

$$\frac{1}{T}\sum_{t=1}^{T}\Delta\widetilde{Z}_{1,t}\widetilde{Z}'_{1,t-1} = (\beta'\alpha)\frac{1}{T}\sum_{t=1}^{T}\widetilde{Z}_{1,t-1}\widetilde{Z}'_{1,t-1} + \frac{1}{T}\sum_{t=1}^{T} v_{1,t}\widetilde{Z}'_{1,t-1}\ .$$

This implies that

$$||\frac{1}{T}\sum_{t=1}^{T}\Delta\widetilde{Z}_{1,t}\widetilde{Z}_{1,t-1} - (\beta'\alpha)\Sigma_{z1.\Delta x}||_F = O_p(\frac{r}{\sqrt{T}}) \tag{3.24}$$

due to (3.23) and since $\frac{1}{T}\sum_{t=1}^{T} v_{1,t}\widetilde{Z}'_{1,t-1} = R_2$ with $||R_1 + R_2||_F = O_P(r/\sqrt{T})$ together with Lemma 3.A.2.

3. mixed stationary/nonstationary block $b_{12} = \frac{1}{T}\Delta Z_1 M Z'_{2,-1}$
From (3.5) we get

$$\widetilde{Z}_{2,t} = \sum_{s=1}^{t}\widetilde{v}_{2,s} = \sum_{s=1}^{t} v_{2,s} - R_3 \tag{3.25}$$

with $||R_3||_F = O_P(r/\sqrt{T})$ since $||\frac{1}{T}\sum_{s=1}^{T} v_{2,s}\Delta X'_{s-1}||_2 = O_P(\frac{1}{\sqrt{T}})$ and Lemma 3.A.2. Thus $\frac{1}{T}\Delta Z_1 M Z'_{2,-1}$ can be further decomposed from (3.5) in ΔZ_1 by summation by part in $Z_{2,-1}$ as:

$$\frac{1}{T}\sum_{t=0}^{T}\Delta\widetilde{Z}_{1,t}\widetilde{Z}'_{2,t-1} \quad = \quad -\frac{1}{T}(\beta'\alpha + I_r)\sum_{t=1}^{T}\widetilde{Z}_{1,t-1}v'_{2,t} - \frac{1}{T}\sum_{t=1}^{T}\widetilde{v}_{1,t}\widetilde{v}'_{2,t} + R_4$$

with $\frac{1}{T}\sum_{t=1}^{T}\widetilde{v}_{1,t}\widetilde{v}'_{2,t} = \Sigma_{v1v2} + R_5$ where $||R_4 + R_5||_F = O_P(\sqrt{mr}/\sqrt{T})$. Hence we get

$$||\frac{1}{T}\sum_{t=0}^{T}\Delta\widetilde{Z}_{1,t}\widetilde{Z}'_{2,t-1} + \Sigma_{v1v2}||_F = O_p(\frac{\sqrt{mr}}{\sqrt{T}}) \tag{3.26}$$

4. mixed stationary/nonstationary $b_{21} = \frac{1}{T}\Delta Z_2 M Z'_{1,-1}$
With (3.25) it holds that $||\frac{1}{T}\sum_{t=1}^{T}\Delta\widetilde{Z}_{2,t}\widetilde{Z}'_{1,t-1}||_F = ||\frac{1}{T}\sum_{t=1}^{T} v_{2,t}\widetilde{Z}'_{1,t-1}||_F + O_p(\frac{\sqrt{mr}}{\sqrt{T}})$ which leads to

$$||\frac{1}{T}\sum_{t=1}^{T}\Delta\widetilde{Z}_{2,t}\widetilde{Z}'_{1,t-1}||_F = O_p(\frac{\sqrt{mr}}{\sqrt{T}}) \tag{3.27}$$

due to the independence condition in Assumption 3.2.1 and Lemma 3.A.2.

5. purely nonstationary block $b_{22} = \frac{1}{T}\Delta Z_2 M Z'_{2,-1}$
From (3.25) we have

$$\frac{1}{T}\sum_{t=1}^{T}\Delta\widetilde{Z}_{2,t}\widetilde{Z}'_{2,t-1} \quad = \quad \frac{1}{T}\sum_{t=1}^{T} v_{2,t}\widetilde{Z}'_{2,t-1} + R_6$$

with $||R_6||_F = O_p(\frac{\sqrt{mr}}{\sqrt{T}})$. We get componentwise for $i, j = 1\ldots, m$ in the leading term on the right that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\widetilde{Z}^i_{2,t-1}v^j_{2,t} - \int_0^1 \mathbf{M}_{2,i}(s)d\mathbf{M}_{2,j}(s) \\
= \quad & \sum_{t=1}^{T}\Big(\frac{1}{\sqrt{T}}\widetilde{Z}^i_{2,t-1}(\frac{1}{\sqrt{T}}v^j_{2,t} - \int_{\frac{t-1}{T}}^{\frac{t}{T}} d\mathbf{M}_{2,j}(s)) + \int_{\frac{t-1}{T}}^{\frac{t}{T}}[\frac{1}{\sqrt{T}}\widetilde{Z}^i_{2,t-1} - \mathbf{M}_{2,i}(s)]d\mathbf{M}_{2,j}(s)\Big) \\
=_d \quad & \sum_{t=1}^{T}\Big(\frac{1}{\sqrt{T}}\widetilde{Z}^i_{2,t-1}(\frac{1}{\sqrt{T}}v^j_{2,t} - \mathbf{M}_{2,j}(\frac{1}{T})) + \int_{\frac{t-1}{T}}^{\frac{t}{T}}[\frac{1}{\sqrt{T}}\widetilde{Z}^i_{2,t-1} - \mathbf{M}_{2,i}(s)]d\mathbf{M}_{2,j}(s)\Big) \qquad (3.28)
\end{aligned}$$

For the first term, define $h^j_t = \frac{1}{\sqrt{T}}v^j_{2,t} - \mathbf{M}_{2,j}(\frac{1}{T})$ and $H^j_n = \sum_{t=1}^{n} h^j_t$. Then by integration by parts,

$$\begin{aligned}
\sum_{t=1}^{T}\frac{1}{\sqrt{T}}\widetilde{Z}^i_{2,t-1}h^j_t \quad &= \quad \sum_{t=1}^{T}\Big(\frac{1}{\sqrt{T}}\widetilde{Z}^i_{2,t-1}(\frac{1}{\sqrt{T}}v^j_{2,t} - \mathbf{M}_{2,j}(\frac{1}{T}))\\
&= \quad \frac{1}{\sqrt{T}}\widetilde{Z}^i_{2,T-1}H^j_T - \frac{1}{\sqrt{T}}\sum_{t=1}^{T-1}\widetilde{v}^i_{2,t}H^j_t \\
&= \quad \frac{1}{\sqrt{T}}\widetilde{Z}^i_{2,T-1}H^j_T - \frac{1}{\sqrt{T}}\sum_{t=1}^{T-1}v^i_{2,t}H^j_t + O_p(\frac{1}{\sqrt{T}})
\end{aligned}$$

Therefore by strong invariance principle (see Theorem 12.7 of DasGupta (2008)), $sup_{t\leqslant T}|H^j_t| = O_p(\frac{(\log T)^{1/2}(\log\log T)^{1/4}}{T^{1/4}})$, which provides an upper bound for variance of the middle term. Therefore, we can conclude that $|\sum_{t=1}^{T}\frac{1}{\sqrt{T}}\widetilde{Z}^i_{2,t-1}h^j_t| = O_p(\frac{(\log T)^{1/2}(\log\log T)^{1/4}}{T^{1/4}})$. Along the same lines it can also be shown that for the second term in (3.28) that

$$\sum_{t=1}^{T}\int_{\frac{t-1}{T}}^{\frac{t}{T}}[\frac{1}{\sqrt{T}}\widetilde{Z}^i_{2,t-1} - \mathbf{M}_{2,i}(s)]d\mathbf{M}_{2,j}(s) = O_p(\frac{(\log T)^{1/2}(\log\log T)^{1/4}}{T^{1/4}}).$$

Therefore we get from (3.28) that $|\frac{1}{T}\sum_{t=1}^{T}\widetilde{Z}^i_{2,t-1}v^j_{2,t} - \int_0^1 \mathbf{M}_{2,i}(s)d\mathbf{M}_{2,j}(s)| = O_p(\frac{(\log T)^{1/2}(\log\log T)^{1/4}}{T^{1/4}})$.

Hence in total, we can conclude that

$$||\frac{1}{T}\sum_{t=1}^{T} v_{2,t}\widetilde{Z}'_{2,t-1} - \int_0^1 d\mathbf{M}_2(s)\mathbf{M}_2(s)'||_F = O_p(\frac{m(\log T)^{1/2}(\log\log T)^{1/4}}{T^{1/4}}) \quad (3.29)$$

Thus from equations (3.24)-(3.29) for the blocks $b_{11}, b_{12}, b_{21}, b_{22}$ we get the first part (3.21) of the initial claim.

6. mixed stationary/nonstationary block $\chi_{12} = \frac{1}{T} Z_{1,-1} M Z'_{2,-1}$
From equation (3.5) we get with negligible R_7 that

$$\frac{1}{T}\sum_{t=1}^{T} \Delta\widetilde{Z}_{1,t}\widetilde{Z}'_{2,t-1} = \frac{1}{T}\sum_{t=1}^{T}(\beta'\alpha)\widetilde{Z}_{1,t-1}\widetilde{Z}'_{2,t-1} + \frac{1}{T}\sum_{t=1}^{T} v_{1,t}\widetilde{Z}'_{2,t-1} + R_7$$

Rearranging yields

$$\frac{1}{T}\sum_{t=1}^{T} \widetilde{Z}_{1,t-1}\widetilde{Z}'_{2,t-1} = (\beta'\alpha)^{-1}(\frac{1}{T}\sum_{t=1}^{T} \Delta\widetilde{Z}_{1,t}\widetilde{Z}'_{2,t-1} - \frac{1}{T}\sum_{t=1}^{T} v_{1,t}\widetilde{Z}'_{2,t-1}) + \bar{R}_7.$$

As the first term on the right has been treated in block 3 above we can use (3.26). For the second term, the standard Brownian motion limit result applies. Moreover, we use that by Assumption 3.2.3 we have $||(\beta'\alpha)^{-1}||_2 = O(r^{\tau_1})$. Hence in total, we find

$$\begin{aligned} & ||\frac{1}{T}\sum_{t=1}^{T} \widetilde{Z}_{1,t-1}\widetilde{Z}'_{2,t-1} + (\beta'\alpha)^{-1}(\Sigma_{v1v2} + \int_0^1 d\mathbf{M}_1\mathbf{M}'_2)||_F \\ = \; & O\Big(r^{\tau_1}\sqrt{\frac{mr}{T} + \frac{mr(\log T)(\log\log T)^{1/2}}{\sqrt{T}}}\Big) \qquad (3.30) \end{aligned}$$

7. mixed stationary/nonstationary block $\chi_{21} = \frac{1}{T}\left(\frac{1}{T} Z_{2,-1} M Z'_{1,-1}\right)$
From χ_{12} in block 6, we know that each element in $\frac{1}{T}\sum_{t=1}^{T}\widetilde{Z}_{1,t-1}\widetilde{Z}'_{2,t-1}$ can at least be bounded to be $O_p(r^{\tau_1})$. This bound is sufficient as for (3.22) the pre-multiplication with D_T^{-1} requires only to study χ_{21} which divides once more by T. Therefore we get similar to (3.30)

$$||\frac{1}{T^2}\sum_{t=1}^{T} \widetilde{Z}_{2,t-1} Z'_{1,t-1}||_F = O_p(\frac{\sqrt{mr}r^{\tau_1}}{T}) \quad (3.31)$$

8. purely non-stationary block $\chi_{22} = \frac{1}{T}\left(\frac{1}{T} Z_{2,-1} M Z'_{2,-1}\right)$

Similar to b_{22} from block 5 but now with an additional T in the denominator from the pre-multiplication of D_T^{-1} in (3.22) we get element-wise, for each $i, j = 1, \dots, m-r$

$$\frac{1}{T^2}\sum_{t=1}^{T} \widetilde{Z}_{2,t-1}^{i}\widetilde{Z}_{2,t-1}^{j} - \int_0^1 \mathbf{M}_{2,i}(s)\mathbf{M}_{2,j}(s)ds = \tag{3.32}$$
$$= \sum_{t=1}^{T}\int_{\frac{t-1}{T}}^{\frac{t}{T}} \frac{1}{\sqrt{T}}\widetilde{Z}_{2,t-1}^{i}\Big(\frac{1}{\sqrt{T}}\widetilde{Z}_{2,t-1}^{j} - \mathbf{M}_{2,j}(s)\Big)ds + \sum_{t=1}^{T}\int_{\frac{t-1}{T}}^{\frac{t}{T}} \Big(\frac{1}{\sqrt{T}}\widetilde{Z}_{2,t-1}^{i} - \mathbf{M}_{2,i}(s)\Big)\mathbf{M}_{2,j}(s)ds$$

From the strong invariance principle for i.i.d. random variables, we have $|\frac{1}{\sqrt{T}}\widetilde{Z}_{2,t-1}^{i} - \mathbf{M}_{2,i}(\frac{t-1}{T})| = O_p(\frac{(\log T)^{1/2}(\log\log T)^{1/4}}{T^{1/4}})$ and for any Brownian motion component in $\mathbf{M}(s)$, $\max_{\frac{t-1}{T}\leqslant s\leqslant\frac{t}{T}} |\mathbf{M}_j(s) - \mathbf{M}_j(\frac{t-1}{T})| = O_p(\sqrt{\frac{\log T}{T}})$ according to Lévy's modulus of continuity theorem. Thus (3.32) is bounded by $O_p(\frac{(\log T)^{1/2}(\log\log T)^{1/4}}{T^{1/4}})$. This yields

$$||\frac{1}{T^2}\sum_{t=1}^{T} \widetilde{Z}_{2,t-1}\widetilde{Z}_{2,t-1}' - \int_0^1 \mathbf{M}_2(s)\mathbf{M}_2'(s)ds||_F = O_p(m\frac{(\log T)^{1/2}(\log\log T)^{1/4}}{T^{1/4}}) \tag{3.33}$$

Combining the blockwise results (3.23),(3.30)–(3.33) for $\chi_{11}, \chi_{12}, \chi_{21}, \chi_{22}$ we get the second part of the initial claim (3.22).

For the final result in ψ, define $\xi = \chi^{-1} = \Big(D_T\frac{1}{T}\sum_{t=1}^{T}\widetilde{Z}_{t-1}\widetilde{Z}_{t-1}'\Big)^{-1}$. Then we get the corresponding blocks of ξ by blockwise inverting as:

$$\begin{aligned}\xi_{11} &= (\chi_{11} - \chi_{12}\chi_{22}\chi_{21})^{-1}\\ \xi_{12} &= -\xi_{11}\chi_{12}\chi_{22}^{-1}\\ \xi_{21} &= -\xi_{22}\chi_{21}\chi_{11}^{-1}\\ \xi_{22} &= (\chi_{22} - \chi_{21}\chi_{11}\chi_{12})^{-1}\end{aligned}$$

Note that any term containing χ_{21} is of smaller order than the others as $||\chi_{21}||_F = O_p(\frac{\sqrt{mr}r^{\tau_1}}{T})$ due to (3.31). Therefore we find with (3.23),(3.30)-(3.33) and Lemma 3.A.1

that

$$\begin{aligned}
&||\xi_{11} - \Sigma^{-1}_{z1.\Delta x}||_F = O_p(\frac{r}{\sqrt{T}})\\
&||\xi_{12} - \Sigma^{-1}_{z1.\Delta x}(\beta'\alpha)^{-1}\big(\Sigma_{v1v2}(\int_0^1 \mathbf{M}_2(s)\mathbf{M}_2'(s)ds)^{-1} + \mathbf{V}_{12})||_F\\
&= O_p(r^{\tau_1}\sqrt{\frac{mr}{T} + \frac{mr(\log T)(\log\log T)^{1/2}}{\sqrt{T}}})\\
&||\xi_{21}||_F = O_p(\frac{\sqrt{mr}r^{\tau_1}}{T})\\
&||\xi_{22} - (\int_0^1 \mathbf{M}_2(s)\mathbf{M}_2'(s)ds)^{-1}||_F = O_p(m\frac{(\log T)^{1/2}(\log\log T)^{1/4}}{T^{1/4}}) \qquad (3.34)
\end{aligned}$$

Thus we get for $\widetilde{\Psi} = (\frac{1}{T}\sum_{t=1}^{T}\Delta\widetilde{Z}_t\widetilde{Z}'_{t-1})\xi$ from (3.21) and (3.34) together with Lemma 3.A.1 and the assumption $r = O(m^{\frac{1}{2\tau_1+1}})$ that

$$\begin{aligned}
||\widetilde{\Psi}_{11} - (\beta'\alpha)||_F &= O_p(\frac{r}{\sqrt{T}})\\
||\widetilde{\Psi}_{12} - \mathbf{V}_{12}||_F &= O_p(m\sqrt{\frac{(\log T)(\log\log T)^{1/2}}{\sqrt{T}}})\\
||\widetilde{\Psi}_{21}||_F &= O_p(\sqrt{\frac{mr}{T}})\\
||\widetilde{\Psi}_{22} - \mathbf{V}_{22}||_F &= O_p(m\sqrt{\frac{(\log T)(\log\log T)^{1/2}}{\sqrt{T}}})
\end{aligned}$$

□

Proof of Corollary 3.2.1

Proof. The proof follows directly from Theorem 3.2.1 with $\Psi_0 = E(\Psi)$ and the weak law of large numbers. □

Proof of Theorem 3.2.2

Proof. Let us first derive two general assertions by which we show that the specific claims of the theorem are implied. Define

$$\beta_0 = \begin{bmatrix} \beta' \\ \beta'_\perp \end{bmatrix}$$

We pre-multiply $\widetilde{\Pi}'$ by matrix β_0. Thus we get with $\widetilde{\Psi} = Q\widetilde{\Pi}Q^{-1}D_T$ as in Theorem 3.2.1

$$\begin{aligned}\beta_0\widetilde{\Pi}' &= \begin{pmatrix} \beta'\widetilde{\Pi}' \\ \beta_\perp'\widetilde{\Pi}' \end{pmatrix} = \begin{pmatrix} I_r & \frac{1}{T}\beta'\alpha_\perp \\ 0 & \frac{1}{T}\beta_\perp'\alpha_\perp \end{pmatrix} \left(Q^{-1}\widetilde{\Psi}\right)' \\ &= \begin{pmatrix} I_r & \frac{1}{T}\beta'\alpha_\perp \\ 0 & \frac{1}{T}\beta_\perp'\alpha_\perp \end{pmatrix} \begin{pmatrix} \alpha(\beta'\alpha)^{-1}\widetilde{\Psi}_{11} + \beta_\perp(\alpha_\perp'\beta_\perp)^{-1}\widetilde{\Psi}_{21} & \alpha(\beta'\alpha)^{-1}\widetilde{\Psi}_{12} + \beta_\perp(\alpha_\perp'\beta_\perp)^{-1}\widetilde{\Psi}_{22} \end{pmatrix}'\end{aligned} \tag{3.35}$$

For the left block of $\left(Q^{-1}\widetilde{\Psi}\right)'$ we use that by Theorem 3.2.1, $||\widetilde{\Psi}_{21}||_F = O_P\left(\sqrt{mr/T}\right)$ and that $||\widetilde{\Psi}_{11} - (\beta'\alpha)||_F = O_P\left(rT^{-1/2}\right)$. Therefore we get from this part for the first block on the left hand side of (3.35) that

$$||\beta'\widetilde{\Pi}' - \alpha'||_F = O_p\left(\frac{r}{\sqrt{T}} + \sqrt{\frac{mr}{T}}\right) = O_P\left(\sqrt{\frac{mr}{T}}\right) \tag{3.36}$$

Note that (3.36) identifies the space of α up to rotation, as we can write without loss of generality $\alpha = \alpha_0\Lambda_0$ with orthonormal α_0 and Λ_0 a diagonal matrix with singular values of α. Then

$$\begin{aligned}||\alpha(\beta'\alpha)^{-1}\widetilde{\Psi}_{11} - \alpha||_F &= ||\alpha_0(\beta'\alpha_0)^{-1}\widetilde{\Psi}_{11} - \alpha_0\Lambda_0||_F \leqslant ||\alpha_0(\beta'\alpha_0)^{-1}||_2||\widetilde{\Psi}_{11} - (\beta'\alpha_0)\Lambda_0||_F = \\ &= ||\alpha_0(\beta'\alpha_0)^{-1}||_2||\widetilde{\Psi}_{11} - (\beta'\alpha)||_F = O_p(\frac{r}{\sqrt{T}}) \ .\end{aligned}$$

For the second block on the left hand side of (3.35), note that $\widetilde{\Psi}_{12}, \widetilde{\Psi}_{22}$ have bounded l_2 norms under our assumptions on m and r. Therefore we get

$$||\beta_\perp'\widetilde{\Pi}'||_2 = O_p(\frac{1}{T}) \ . \tag{3.37}$$

We now use (3.36) and (3.37) in order to prove the stated claims of the theorem in reverse order and start with part 2. Due to the unitary invariance property of singular values, we have

$$\sigma_j(\beta_0\widetilde{\Pi}') = \sigma_j(S\widetilde{\Pi}') = \sigma_j(\widetilde{R}) \tag{3.38}$$

for all $j = 1, \ldots m$. With equation (3.36), this implies in particular that

$$|\sigma_j(\widetilde{R}) - \sigma_j(\alpha)| = O_p(\sqrt{\frac{mr}{T}}) \quad \text{for } j = 1, \ldots, r \tag{3.39}$$

due to matrix perturbation theory (Mirsky version, Theorem 4.11 of Stewart and Sun (1990)).

The column-pivoting step in the QR decomposition makes the $\widetilde{R}_{11}$ a well-conditioned matrix, thus the largest r singular values in $\widetilde{R}$ are in $\widetilde{R}_1$ which contains the first r-

rows. Besides, the strict upper-triangular structure of $\widetilde{R}$ excludes linear dependence between any two rows in $\widetilde{R}_1$. Therefore, we can conclude that

$$\sigma_r(\widetilde{R}) \leqslant \sqrt{\sum_{j=k}^{m} \widetilde{R}(k,j)^2} \leqslant \sigma_1(\widetilde{R}) \quad \text{for } k = 1, \ldots, r \tag{3.40}$$

The matrix perturbation theory result (3.39) provides further bounds for l_2 norm of each row in $\widetilde{R}_1$, i.e.,

$$\sigma_r(\widetilde{R}) \geqslant \sigma_r(\alpha) - O_p(\sqrt{\frac{mr}{T}})$$
$$\sigma_1(\widetilde{R}) \leqslant \sigma_1(\alpha) + O_p(\sqrt{\frac{mr}{T}})$$

In the same way we obtain from (3.38) together with (3.37), that

$$|\sigma_j(\widetilde{R})| = O_p(1/T) \tag{3.41}$$

for $j = r+1, \ldots, m$. With the upper triangular structure column pivoting in $\widetilde{R}$, this implies $\sqrt{\sum_{j=k}^{m} \widetilde{R}(k,j)^2} = O_p(1/T)$ for $k = r+1, \ldots, m$ Thus we have shown claim 2 of the theorem.

Moreover, (3.41) implies that $||\widetilde{R}_{22}||_F = O_p(\frac{\sqrt{m}}{T})$. We can generate a square matrix $\widetilde{R}_1^0$ by adding $m-r$ rows of zeros to $\widetilde{R}_1$. Then $\sigma_j(\widetilde{R}_1^0) = \sigma_j(\widetilde{R}_1)$ for $j \leqslant r$ and $\sigma_j(\widetilde{R}_1^0) = 0$ if $j > r$. Therefore, by the fact that $||\widetilde{R} - \widetilde{R}_1^0||_F = ||\widetilde{R}_{22}||_F$, we can conclude that

$$|\sigma_j(\widetilde{R}) - \sigma_j(\widetilde{R}_1)| = O_p(\frac{\sqrt{m}}{T}), \qquad j = 1, .., r$$

and thus

$$|\sigma_j(\widetilde{R}_1) - \sigma_j(\alpha)| = O_p(\sqrt{\frac{mr}{T}}) \quad \text{for } j = 1, \ldots, r \tag{3.42}$$

Thus we have shown claim 3 of the theorem.

In order to show part 1 of the theorem, we re-write $\beta_0\widetilde{\Pi}'$ with the QR-decomposition components of $\widetilde{\Pi}$ as follows

$$\begin{pmatrix} \beta'\widetilde{\Pi}' \\ \beta_\perp'\widetilde{\Pi}' \end{pmatrix} = \begin{pmatrix} \beta'\widetilde{S}_1\widetilde{R}_{11} & \beta'\widetilde{S}_1\widetilde{R}_{12} + \beta'\widetilde{S}_2\widetilde{R}_{22} \\ \beta_\perp'\widetilde{S}_1\widetilde{R}_{11} & \beta_\perp'\widetilde{S}_1\widetilde{R}_{12} + \beta_\perp'\widetilde{S}_2\widetilde{R}_{22} \end{pmatrix}. \tag{3.43}$$

By equating (3.35) and (3.43) we get

$$\left(\begin{array}{cc} \beta_\perp' \widetilde{S}_1 \widetilde{R}_{11} & \beta_\perp' \widetilde{S}_1 \widetilde{R}_{12} + \beta_\perp' \widetilde{S}_2 \widetilde{R}_{22} \end{array} \right) = \frac{1}{T}(\beta_\perp' \alpha_\perp)\Big(\alpha(\beta'\alpha)^{-1}\widetilde{\Psi}_{12} + \beta_\perp(\alpha_\perp'\beta_\perp)^{-1}\widetilde{\Psi}_{22}\Big)'$$

which is equivalent to

$$\begin{aligned} \beta_\perp' \widetilde{S}_1 &= -\left[\begin{array}{cc} 0 & \beta_\perp' \widetilde{S}_2 \widetilde{R}_{22} \end{array} \right] \widetilde{R}_1' (\widetilde{R}_1 \widetilde{R}_1')^{-1} \\ &+ \frac{1}{T}(\beta_\perp' \alpha_\perp)\Big(\alpha(\beta'\alpha)^{-1}\widetilde{\Psi}_{12} + \beta_\perp(\alpha_\perp'\beta_\perp)^{-1}\widetilde{\Psi}_{22}\Big)' \widetilde{R}_1' (\widetilde{R}_1 \widetilde{R}_1')^{-1} \quad (3.44) \end{aligned}$$

Note that for the first term on the right hand side of (3.44) we get due to (3.37) that $||\beta_\perp' \widetilde{S}_2 \widetilde{R}_{22}||_2 = O_P(1/T)$. For the second one, Ψ_{12} and Ψ_{22} have bounded l_2 norms since they can be approximated with $\mathbf{V}_{12}$ and $\mathbf{V}_{22}$ due to Theorem 3.2.1 and with Lemma 3.A.1, both $\mathbf{V}_{12}$ and $\mathbf{V}_{22}$ have bounded l_2 norm. Therefore, the upper-bound in l_2 norm for $\beta_\perp' \widetilde{S}_1$ is driven by the rate of $(\widetilde{R}_1 \widetilde{R}_1')^{-1}/T$. From (3.42) we have that the singular values of $\widetilde{R}_1$ can be approximated by those of α. Therefore using Assumption 3.2.3 we conclude in total that $||\beta_\perp' \widetilde{S}_1||_F = O_p(\frac{\sqrt{m}r^{2\tau_1}}{T})$.

□

Proof of Theorem 3.2.3

Proof. The main idea of the proof is to show that the group-wise KKT condition holds with high probability. The assumptions on λ_T^{rank} ensure that the penalty on the stationary cointegrated part decays to zero while that on the unit root part diverges fulfilling the irrepresentible condition proposed in Zhao and Yu (2006).

Denote by $\widetilde{S}'\widetilde{Y}_{t-1} = \left[\begin{array}{c} \breve{Z}_{1,t-1} \\ \breve{Z}_{2,t-1} \end{array} \right]$ where $\breve{Z}_{1,t-1}$ is the projection of Y_{t-1} onto the subspace generated by $\widetilde{S}_1$. According to Theorem 3.2.2, the subspace distance between $\widetilde{S}_1$ and β , i.e. $||\widetilde{S}_1 - \beta H||_F$ for some orthonormal H, converges at a faster rate than the subspace distance of $\widetilde{R}_1$ and α under the given conditions on m and r. Therefore the first step estimation error from using $\widetilde{S}$ in (3.8) instead of the infeasible true S_1 is negligible and wlog. we use $\widetilde{Z}_{1,t-1}$ instead of $\breve{Z}_{1,t-1}$ and $\widetilde{Z}_{2,t-1}$ instead of $\breve{Z}_{2,t-1}$ for the rest of this proof for ease of notation.

Since α and β are only identified up to rotation, we write wlog $\bar{\alpha} = \alpha H$ with H as defined for $\widetilde{S}_1$. Note that $\bar{\alpha}$ and α describe the same space. Define $\bar{\alpha}_0 = [\bar{\alpha}, 0_{m\times m-r}]$,

$\delta_R = \hat{R}' - \bar{\alpha}_0$ and δ_{R_1} for the first r columns in δ_R. Then we have

$$\begin{aligned}
& \sum_{t=1}^{T} \parallel \Delta\widetilde{Y}_t - (\widetilde{Z}'_{t-1} \otimes I_m) vec(\hat{R}') \parallel^2 + \sum_{j=1}^{m} \frac{\lambda_T^{rank}}{\tilde{\mu}_j^{\gamma}} ||\hat{R}'(,j)||_2 \\
= & \sum_{t=1}^{T} \parallel \tilde{w}_t - (\widetilde{Z}'_{t-1} \otimes I_m) vec(\delta_R) \parallel^2 + \sum_{j=1}^{m} \frac{\lambda_T^{rank}}{\tilde{\mu}_j^{\gamma}} ||\bar{\alpha}_0(,j) + \delta_R(,j)||_2 \\
= & \sum_{t=1}^{T} \tilde{w}'_t \tilde{w}_t - 2w'_t(\widetilde{Z}'_{t-1} \otimes I_m) vec(\delta_R) + vec(\delta_R)'(\widetilde{Z}_{t-1}\widetilde{Z}'_{t-1} \otimes I_m) vec(\delta_R) \\
& + \sum_{j=1}^{m} \frac{\lambda_T^{rank}}{\tilde{\mu}_j^{\gamma}} ||\bar{\alpha}_0(,j) + \delta_R(,j)||_2
\end{aligned}$$

Therefore, the minimization of (3.8) in $\hat{R}$ is equivalent to minimizing

$$\sum_{t=1}^{T} -2w'_t(\widetilde{Z}'_{t-1} \otimes I_m) vec(\delta_R) + vec(\delta_R)'(\widetilde{Z}_{t-1}\widetilde{Z}'_{t-1} \otimes I_m) vec(\delta_R) + \sum_{j=1}^{m} \frac{\lambda_T^{rank}}{\tilde{\mu}_j^{\gamma}} ||\bar{\alpha}_0(,j) + \delta_R(,j)||_2 \tag{3.45}$$

in δ_R. With $D_{1T} = diag\{\sqrt{T} I_r, T I_{m-r}\}$ the term inside the first sum can be written as
$-2w'_t(\widetilde{Z}'_{t-1} D_{1T}^{-1} \otimes I_m) vec(\delta_R D_{1T}) + vec(\delta_R D_{1T})'(D_{1T}^{-1} \widetilde{Z}_{t-1} \widetilde{Z}'_{t-1} D_{1T}^{-1} \otimes I_m) vec(\delta_R D_{1T})$.
Thus the Karush-Kuhn-Tucker (KKT) condition for group-wise variable selection from (3.45) is

$$-\frac{1}{\sqrt{T}} \sum_{t=1}^{T} w_t \widetilde{Z}'_{1,t-1} + \sqrt{T} \delta_{R1} \frac{1}{T} \sum_{t=1}^{T} \widetilde{Z}_{1,t-1} \widetilde{Z}'_{1,t-1} = -[\frac{\bar{\lambda}_{1,T}}{2\sqrt{T}} \frac{\bar{\alpha}(,1)}{||\bar{\alpha}(,1)||_2}, \ldots, \frac{\bar{\lambda}_{r,T}}{2\sqrt{T}} \frac{\bar{\alpha}(,r)}{||\bar{\alpha}(,r)||_2}] \tag{3.46}$$

$$||\Big(\sum_{t=1}^{T} -2\frac{1}{T} w_t \widetilde{Z}'_{2,t-1} + 2\sqrt{T} \delta_{R1} \frac{1}{T^{3/2}} \widetilde{Z}_{1,t-1} \widetilde{Z}'_{2,t-1}\Big)_j ||_2 < \frac{\bar{\lambda}_{r+j,T}}{T} \tag{3.47}$$

where $\bar{\lambda}_{j,T} = \frac{\lambda_T^{rank}}{\tilde{\mu}_j^{\gamma}}$ and the subscript j denotes the jth column. The expression follows since the derivative of the first term in (3.45) w.r.t. $vec(\delta_R D_{1T})$ is $\sum_{t=1}^{T} -2(D_{1T}^{-1} \widetilde{Z}_{t-1} \otimes I_m) w_t + 2(D_{1T}^{-1} \widetilde{Z}_{t-1} \widetilde{Z}'_{t-1} D_{1T}^{-1} \otimes I_m) vec(\delta_R D_{1T}) = \sum_{t=1}^{T} -2w_t \widetilde{Z}'_{t-1} D_{1T}^{-1} + 2\delta_R D_{1T} D_{1T}^{-1} \widetilde{Z}_{t-1} \widetilde{Z}'_{t-1} D_{1T}^{-1}$.

Define $V_\alpha = [\frac{\bar{\alpha}(,1)}{||\bar{\alpha}(,1)||_2 \tilde{\mu}_1^{\gamma}}, \ldots, \frac{\bar{\alpha}(,r)}{||\bar{\alpha}(,r)||_2 \tilde{\mu}_r^{\gamma}}]$ and

$$\begin{aligned}
S_{z1z1} &= \frac{1}{T} \sum_{t=1}^{T} \widetilde{Z}_{1,t-1} \widetilde{Z}'_{1,t-1} & S_{wz1} &= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} w_t \widetilde{Z}'_{1,t-1} \\
S_{z1z2} &= \frac{1}{T^{3/2}} \sum_{t=1}^{T} \widetilde{Z}_{1,t-1} \widetilde{Z}'_{2,t-1} & S_{wz2} &= \frac{1}{T} \sum_{t=1}^{T} w_t \widetilde{Z}'_{2,t-1}
\end{aligned}$$

From the proof of Theorem 3.2.1, we can use that $S_{z1z2} = \chi_{12}/\sqrt{T}$ in block 6.

Thus we can conclude from (3.30) that $||S_{z1z2}||_2 = O_p(\frac{r^{\tau_1}}{\sqrt{T}})$. Moreover, (3.23) and Lemma 3.A.2 imply that S_{z1z1} and S_{wz1} have bounded l_2 norm. Therefore we can re-write the first KKT-conditions (3.46) for the stationary part as

$$\sqrt{T}\delta_{R1} \quad = \quad -\frac{\lambda_T^{rank}}{2\sqrt{T}} V_\alpha S_{z1z1}^{-1} + S_{wz1}S_{z1z1}^{-1} \tag{3.48}$$

which implies that

$$||\sqrt{T}\delta_{R1}||_2 \quad = \quad O_p(\frac{\lambda_T^{rank}}{\sqrt{T}} r^{\tau_1\gamma+\frac{1}{2}} + 1) = o_P(1) \tag{3.49}$$

The convergence in (3.49) follows from the condition on the tuning parameter $\frac{\lambda_T^{rank}}{\sqrt{T}} r^{\tau_1\gamma+\frac{1}{2}} \to 0$ in the theorem. It thus yields that each element in δ_{R1} converges to zero at the rate of $\sqrt{T}$. Hence, the first r columns of the solution $\hat{R}'$ in (3.8) are $\sqrt{T}$-consistent for $\bar{\alpha}$.

Moreover, for the second part (3.47) of the KKT conditions, we plug in (3.48). Hence for the exclusion of the non-stationary components from (3.8), it is sufficient if

$$||(S_{wz1}S_{z1z1}^{-1}S_{z1z2} - S_{wz2})_k||_2 < \frac{\lambda_T^{rank}}{2T}\tilde{\mu}_{r+k}^{-\gamma} - \frac{\lambda_T^{rank}}{2\sqrt{T}}||(V_\alpha S_{z1z1}^{-1}S_{z1z2})_k||_2 \tag{3.50}$$

for $k = 1, 2, \ldots, m - r$. It remains to show that (3.50) is bounded in probability which implies selection consistency holds with probability one.

$$\begin{aligned}
\frac{\lambda_T^{rank}}{\sqrt{T}}||(V_\alpha S_{z1z1}^{-1}S_{z1z2})_k||_2 \quad &\leqslant \quad \frac{\lambda_T^{rank}}{\sqrt{T}}||V_\alpha S_{z1z1}^{-1}S_{z1z2}||_F \\
&\leqslant \quad \frac{\lambda_T^{rank}}{\sqrt{T}}||V_\alpha||_F||S_{z1z1}^{-1}||_2||S_{z1z2}||_2 \\
&= \quad O_p(\frac{\lambda_T^{rank}}{\sqrt{T}} r^{\tau_1\gamma+1/2}\frac{r^{\tau_1}}{\sqrt{T}})
\end{aligned}$$

Thus the RHS of (3.50) is dominated by the first term. The LHS of (3.50) is dominated by S_{wz2} since $||S_{wz1}S_{z1z1}^{-1}S_{z1z2}||_2 = O_P(r^{\tau_1}/\sqrt{T})$ due to (3.23), (3.30), (3.26) and Lemma 3.A.2.

Moreover, for S_{wz2} we use that $\bar{Z}_{2,t} = \sum_{s=1}^t \alpha'_\perp w_s$ as in (3.25) to get for all $i =$

$1, 2, \ldots, m$ and $j = 1, 2, \ldots, m-r$

$$\begin{aligned}
\mathbf{E}(\frac{1}{T}\sum_{t=1}^{T} w_t^i \widetilde{Z}_{2,t-1}^j)^2 &= \frac{1}{T^2}\mathbf{E}(\sum_{t=1}^{T}(w_t^i)^2(\bar{Z}_{2,t-1}^j)^2) + \frac{1}{T^2}\mathbf{E}(\sum_{s\neq t}^{T} w_s^i \bar{Z}_{2,s-1}^j w_t^i \bar{Z}_{2,t-1}^j) + o_p(1) \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\mathbf{E}((w_t^i)^2)\mathbf{E}((\bar{Z}_{2,t-1}^j)^2) + o_p(1) = O_p(\frac{1}{T^2}\sum_{t=1}^{T} t) = O_p(1) \qquad (3.51)
\end{aligned}$$

the residual denoted as $o_p(1)$ is due to the difference between $\widetilde{Z}_t$ and $\bar{Z}_t$.

Then we find that with $N_i = (S_{wz2})_{ji}$ for any j , we have that $\left\{\sum_{k=1}^{m} N_k \leqslant c\right\} \supseteq \bigcap_k \left\{N_k \leqslant \frac{c}{m}\right\}$ implies $\left\{\sum_{k=1}^{m} N_k > c\right\} \subseteq \bigcup_k \left\{N_k > \frac{c}{m}\right\}$. Thus we can conclude that

$$\begin{aligned}
\mathbb{P}(\sqrt{\sum_{i=1}^{m} N_i^2} > \frac{\lambda_T^{rank}}{2T}\tilde{\mu}_{r+k}^{-\gamma}) &\leqslant \mathbb{P}(\sum_{i=1}^{m} N_i^2 > \left(\frac{\lambda_T^{rank}}{2T}\tilde{\mu}_{r+k}^{-\gamma}\right)^2) \\
&\leqslant \sum_{i=1}^{m}\mathbb{P}(|N_i| > \frac{\lambda_T^{rank}}{2T\sqrt{m}}\tilde{\mu}_{r+k}^{-\gamma}) \\
&\leqslant mC_0^2(\frac{\lambda_T^{rank}T^{\gamma-1}}{\sqrt{m}})^{-2} \leqslant (\frac{mC_0}{\lambda_T^{rank}T^{\gamma-1}})^2
\end{aligned}$$

for some $0 < C_0 < \infty$ where we use Chebyshev's inequality and (3.51) together with $\widetilde{\mu}_{r+k} = O_P(1/T)$ from Theorem 3.2.2 in the last line. Thus with $\frac{m^{3/2}}{\lambda_T^{rank}T^{\gamma-1}} \to 0$ we simultaneously exclude the last $m-r$ columns with probability tending to 1. □

Lemma 3.A.3. *Let the assumptions of Theorem 3.3.1 hold and* $\breve{Y}, \breve{X}$, *and* $\breve{w}$ *are as defined in* (3.9). *Then*

$$\begin{aligned}
&||\frac{1}{T}\sum_{t=1}^{T}\Delta\breve{Y}_t\Delta\breve{Y}_t' - \Sigma_{\Delta y.z1}||_F = O_p(\frac{m}{\sqrt{T}}) \\
&||\frac{1}{T}\sum_{t=1}^{T}\Delta\breve{X}_{t-1}\Delta\breve{X}_{t-1}' - \Sigma_{\Delta x.z1}||_F = O_p(\frac{m}{\sqrt{T}}) \\
&||\frac{1}{T}\sum_{t=1}^{T}\breve{w}_t\breve{w}_t' - \Sigma_w||_F = O_p(\frac{m}{\sqrt{T}})
\end{aligned}$$

Proof. Let $D_{0T} = diag\{\sqrt{T}I_r, TI_{m-r}\}$ and the matrix Q as defined for (3.5). Then we can write the transformation for lag selection C from (3.9) as

$$\begin{aligned} C &= I_T - Y'_{-1}(Y_{-1}Y'_{-1})^{-1}Y_{-1} = I_T - Y'_{-1}Q'D_{0T}^{-1}(D_{0T}^{-1}QY_{-1}Y'_{-1}Q'D_{0T}^{-1})^{-1}D_{0T}^{-1}QY_{-1} \\ &= I_T - Z'_{-1}D_{0T}^{-1}(D_{0T}^{-1}Z_{-1}Z'_{-1}D_{0T}^{-1})^{-1}D_{0T}^{-1}Z_{-1} \end{aligned}$$

We therefore obtain

$$\Delta\breve{Y}_t = \Delta Y_t - (\textstyle\sum_{t=1}^T \Delta Y_t Z'_{t-1}D_{0T}^{-1})(D_{0T}^{-1}Z_{-1}Z'_{-1}D_{0T}^{-1})^{-1}D_{0T}^{-1}Z_{t-1}$$

$$\Delta\breve{X}_{t-1} = \Delta X_t - (\textstyle\sum_{t=1}^T \Delta X_{t-1} Z'_{t-1}D_{0T}^{-1})(D_{0T}^{-1}Z_{-1}Z'_{-1}D_{0T}^{-1})^{-1}D_{0T}^{-1}Z_{t-1}$$

$$\breve{w}_t = w_t - (\textstyle\sum_{t=1}^T w_t Z'_{t-1}D_{0T}^{-1})(D_{0T}^{-1}Z_{-1}Z'_{-1}D_{0T}^{-1})^{-1}D_{0T}^{-1}Z_{t-1}$$

Denote $S^z = D_{0T}^{-1}Z_{-1}Z'_{-1}D_{0T}^{-1}$ and

$$\begin{aligned} S^z_{11} &= \frac{1}{T}\sum_{t=1}^T Z_{1,t-1}Z'_{1,t-1} \\ S^z_{12} &= \frac{1}{T^{3/2}}\sum_{t=1}^T Z_{1,t-1}Z'_{2,t-1} \\ S^z_{22} &= \frac{1}{T^2}\sum_{t=1}^T Z_{2,t-1}Z'_{2,t-1} \end{aligned}$$

Then $||S^z_{11}||_2 = O_P(1)$ by Lemma 3.A.2. With χ_{12} of block 6 in the proof of Theorem 3.2.1, we have $||S^z_{12}||_2 = ||\chi_{12}/\sqrt{T}||_2 + o_P(1)$ due to Lemma 2. Hence (3.30) implies that $||S^z_{12}||_2 = O_p(\frac{r^{\tau_1}}{\sqrt{T}})$. In the same way, Lemma 3.A.2 and (3.33) yield $||S^z_{22}||_2 = O_p(1)$.

The inverse $S^{z,-1}$ of S^z has the following blockwise form

$$S^{z,-1} = \begin{bmatrix} (S^z_{11} - S^z_{12}S^{z,-1}_{22}S^z_{21})^{-1} & -(S^z_{11} - S^z_{12}S^{z,-1}_{22}S^z_{21})^{-1}S^z_{12}S^{z,-1}_{22} \\ -(S^z_{22} - S^z_{21}S^{z,-1}_{11}S^z_{12})^{-1}S^z_{21}S^{z,-1}_{11} & (S^z_{22} - S^z_{21}S^{z,-1}_{11}S^z_{12})^{-1} \end{bmatrix} \tag{3.52}$$

where $||S^z_{ij}S^{z,-1}_{jj}S^z_{ji}||_2 \leqslant ||S^z_{ij}||_2||S^{z,-1}_{jj}||_2||S^z_{ji}||_2 = O_p(\frac{r^{2\tau_1}}{T})$ for $1 \leqslant i \neq j \leqslant 2$ by the considerations above. We get

$$\begin{aligned} ||S^{z,-1}_{11} - \Sigma^{-1}_{z1}||_F &= O_p(\frac{r}{\sqrt{T}}) \\ ||S^{z,-1}_{12}||_2 &= O_p(\frac{r^{\tau_1}}{\sqrt{T}}) \end{aligned}$$

where the first equation is analogous to (3.23) and the second follows from above.

Then we get

$$
\begin{aligned}
&(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Delta Y_t Z'_{t-1}D_{0T}^{-1})(D_{0T}^{-1}Z_{-1}Z'_{-1}D_{0T}^{-1})^{-1}(\frac{1}{\sqrt{T}}D_{0T}^{-1}\sum_{t=1}^{T}Z_{t-1}\Delta Y'_t)\\
= &\ \frac{1}{T}\sum_{t=1}^{T}\Delta Y_t Z'_{1,t-1}S_{11}^{z,-1}\frac{1}{T}\sum_{t=1}^{T}Z_{1,t-1}\Delta Y'_t + \frac{1}{T^{3/2}}\sum_{t=1}^{T}\Delta Y_t Z'_{2,t-1}S_{21}^{z,-1}\frac{1}{T}\sum_{t=1}^{T}Z_{1,t-1}\Delta Y'_t\\
&+\frac{1}{T}\sum_{t=1}^{T}\Delta Y_t Z'_{1,t-1}S_{21}^{z,-1}\frac{1}{T^{3/2}}\sum_{t=1}^{T}Z_{2,t-1}\Delta Y'_t + \frac{1}{T^{3/2}}\sum_{t=1}^{T}\Delta Y_t Z'_{2,t-1}S_{22}^{z,-1}\frac{1}{T^{3/2}}\sum_{t=1}^{T}Z_{2,t-1}\Delta Y'_t
\end{aligned}
$$

For the first term we get

$$
||\frac{1}{T}\sum_{t=1}^{T}\Delta Y_t Z'_{1,t-1}S_{11}^{z,-1}\frac{1}{T}\sum_{t=1}^{T}Z_{1,t-1}\Delta Y'_t - \Sigma_{\Delta y.z1}||_F = O_p(\frac{m}{\sqrt{T}})
$$

where $\Sigma_{\Delta y.z1} = \mathbf{E}(\Delta Y_t Z'_{1,t-1})\mathbf{E}(Z_{1,t-1}Z'_{1,t-1})^{-1}\mathbf{E}(Z_{1,t-1}\Delta Y'_t)$. For the second one, we have

$$
\frac{1}{T^{3/2}}\sum_{t=1}^{T}\Delta Y_t Z'_{2,t-1} = \frac{1}{T^{3/2}}\sum_{t=1}^{T}\left(\alpha Z_{1,t-1}Z'_{2,t-1} + B\Delta X_{t-1}Z'_{2,t-1} + w_t Z'_{2,t-1}\right),
$$

which implies that $||\frac{1}{T^{3/2}}\sum_{t=1}^{T}\Delta Y_t Z'_{2,t-1}||_2 = O_p(\frac{r^{\tau_1}}{\sqrt{T}})$, as well as $||\frac{1}{T^{3/2}}\sum_{t=1}^{T}Z_{1,t-1}Z'_{2,t-1}||_2 = O_p(\frac{r^{\tau_1}}{\sqrt{T}})$ due to (3.30) and Lemma 2 from the first part of the expression. The l_2 norms for the other two terms are negligible with faster rate $O_P(1/\sqrt{T})$ and Lemma 2. Therefore,

$$
\begin{aligned}
&||\frac{1}{T^{3/2}}\sum_{t=1}^{T}\Delta Y_t Z'_{2,t-1}S_{21}^{z,-1}\frac{1}{T}\sum_{t=1}^{T}Z_{1,t-1}\Delta Y'_t||_F\\
\leqslant &\ ||\frac{1}{T^{3/2}}\sum_{t=1}^{T}\Delta Y_t Z'_{2,t-1}||_2||S_{21}^{z,-1}||_2||\frac{1}{T}\sum_{t=1}^{T}Z_{1,t-1}\Delta Y'_t||_2\sqrt{r}\\
= &\ O_p(\frac{r^{2\tau_1+1/2}}{T}).
\end{aligned}
$$

Thus in total,

$$
||(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Delta Y_t Z'_{t-1}D_{0T}^{-1})(D_{0T}^{-1}Z_{-1}Z'_{-1}D_{0T}^{-1})^{-1}(\frac{1}{\sqrt{T}}D_{0T}^{-1}\sum_{t=1}^{T}Z_{t-1}\Delta Y'_t) - \Sigma_{\Delta y.z1}||_F = O_p(\frac{m}{\sqrt{T}})
$$

Therefore, we conclude that $||\frac{1}{T}\sum_{t=1}^{T}\Delta\breve{Y}_t\Delta\breve{Y}'_t - \Sigma_{\Delta y.z1}||_F = O_p(\frac{m}{\sqrt{T}})$. In the same way, the results for $\Delta\breve{X}_{t-1}$ are obtained. The performance of $\sum_{t=1}^{T}\breve{w}_t\breve{w}'_t$ can be directly inferred from Lemma 3.A.2. □

Proof of Theorem 3.3.1

Proof. For the least squares estimate $\breve{B}$, we consider

$$\sqrt{T}(\breve{B} - B) = (\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\breve{w}_t\Delta\breve{X}'_{t-1})(\frac{1}{T}\sum_{t=1}^{T}\Delta\breve{X}_{t-1}\Delta\breve{X}'_{t-1})^{-1} .$$

Hence we can write the first component with $S^{z,-1}$ from (3.52) explicitly as

$$\begin{aligned}&\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\breve{w}_t\Delta\breve{X}'_{t-1}\\ =\quad & \frac{1}{\sqrt{T}}\sum_{t=1}^{T}w_t\Delta X'_{t-1} - [\frac{1}{\sqrt{T}}\sum_{t=1}^{T}w_tZ'_{1,t-1}, \frac{1}{T}\sum_{t=1}^{T}w_tZ'_{2,t-1}]S^{z,-1}\begin{bmatrix}\frac{1}{T}\sum_{t=1}^{T}Z_{1,t-1}\Delta X'_{t-1}\\ \frac{1}{T^{3/2}}\sum_{t=1}^{T}Z_{2,t-1}\Delta X'_{t-1}\end{bmatrix} .\end{aligned}$$

Thus Lemma 3.A.3 implies that

$$||\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\breve{w}_t\Delta\breve{X}'_{t-1} - \frac{1}{\sqrt{T}}\sum_{t=1}^{T}w_t(\Delta X'_{t-1} - Z'_{1,t-1}\Sigma_{z1}^{-1}\Sigma_{z1\Delta x})||_F = O_p(\frac{m}{\sqrt{T}})$$

and therefore by Lemma 3.A.3 and Lemma 3.A.1 for $(\frac{1}{T}\sum_{t=1}^{T}\Delta\breve{X}_{t-1}\Delta\breve{X}'_{t-1})^{-1}$ it holds that

$$||\sqrt{T}(\breve{B} - B) - \frac{1}{\sqrt{T}}\sum_{t=1}^{T}w_t(\Delta X'_{t-1} - Z'_{1,t-1}\Sigma_{z1}^{-1}\Sigma_{z1\Delta x})\Sigma_{\Delta x.z1}^{-1}||_F = O_p(\frac{m}{\sqrt{T}})$$

With $\Delta\dot{X}_{t-1} = \Sigma_{\Delta x.z1}^{-1}(\Delta X_{t-1} - \Sigma_{\Delta xz1}\Sigma_{z1}^{-1}Z_{1,t-1})$, we can thus conclude that

$$||vec(\breve{B} - B) - vec(\frac{1}{T}\sum_{t=1}^{T}w_t\Delta\dot{X}'_{t-1})||_\infty \leqslant ||(\breve{B} - B) - \frac{1}{T}\sum_{t=1}^{T}w_t\Delta\dot{X}'_{t-1}||_F = O_p(\frac{m}{T^{3/2}})\text{(3.53)}$$

With triangle inequality and the maximal inequality in Chernozhukov et al. (2013), which implies for $vec(w_t\Delta\dot{X}'_{t-1})$

$$||\frac{1}{T}\sum_{t=1}^{T}vec(w_t\Delta\dot{X}'_{t-1})||_\infty = O_p(\sqrt{\frac{\log m}{T}}) , \tag{3.54}$$

this implies the first claim of the theorem. As $\widetilde{B}$ satisfies

$$\begin{aligned}\sqrt{T}(\widetilde{B}-B) &= -\frac{\lambda_T^{ridge}}{\sqrt{T}}B(\frac{1}{T}\sum_{t=1}^{T}\Delta\breve{X}_{t-1}\Delta\breve{X}'_{t-1}+\frac{\lambda_T^{ridge}}{T}I_{mP})^{-1}\\ &\quad +(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}w_t\Delta\breve{X}'_{t-1})(\frac{1}{T}\sum_{t=1}^{T}\Delta\breve{X}_{t-1}\Delta\breve{X}'_{t-1}+\frac{\lambda_T^{ridge}}{T}I_{mP})^{-1}\ ,\end{aligned}$$

for $\frac{\lambda_T^{ridge}}{\sqrt{T}}\to 0$, the asymptotics of $\widetilde{B}$ and $\breve{B}$ coincide. □

Proof of Theorem 3.3.2

Proof. With the Lasso-estimator $\hat{B}$ from (3.12) for $(B_1, B_2, \ldots, B_P)$ define $\delta_B = \hat{B}-B$. Then we get for the first part of the Lasso criterion function (3.12) that

$$\begin{aligned}&\sum_{t=1}^{T}||\Delta\breve{Y}_t-\sum_{j=1}^{P}B_j\Delta\breve{Y}_{t-j}||^2\\ =\ &\sum_{t=1}^{T}(B\Delta\breve{X}_{t-1}+\breve{w}_t-\hat{B}\Delta\breve{X}_{t-1})'(B\Delta\breve{X}_{t-1}+\breve{w}_t-\hat{B}\Delta\breve{X}_{t-1})\\ =\ &\sum_{t=1}^{T}(\breve{w}_t-\delta_B\breve{X}_{t-1})'(\breve{w}_t-\delta_B\breve{X}_{t-1})\\ =\ &\sum_{t=1}^{T}\breve{w}'_t\breve{w}_t-2\frac{1}{\sqrt{T}}w'_t(\breve{X}'_{t-1}\otimes I_m)vec(\sqrt{T}\delta_B)+\frac{1}{T}vec(\sqrt{T}\delta_B)'(\breve{X}_{t-1}\breve{X}'_{t-1}\otimes I_m)vec(\sqrt{T}\delta_B)\ .\end{aligned}$$

Taking first order conditions w.r.t. $vec(\sqrt{T}\delta_B)$ yields

$$\sum_{t=1}^{T}-\frac{2}{\sqrt{T}}(\Delta\breve{X}_{t-1}\otimes I_m)w'_t+\frac{2}{T}(\Delta\breve{X}_{t-1}\Delta\breve{X}'_{t-1}\otimes I_m)vec(\sqrt{T}\delta_B)$$

For the rest of the proof, we assume for ease of notation that all $B_1, B_2, \ldots, B_p$ are non-zero not just B_p as Assumption 3.2.3 implies. Denote by δ_B^0 the first mp columns of δ_B. Consistent lag selection requires that each $m\times m$ block in δ_B^0 contains a non-zero element but the last $m(P-p)$ columns are zero, which can be ensured by the following KKT conditions:

$$\begin{aligned} \sum_{t=1}^{T} -\frac{1}{\sqrt{T}} w_t \Delta\check{X}^{0\prime}_{t-1} + \frac{1}{T}(\sqrt{T}\delta^0_B)(\Delta\check{X}^0_{t-1}\Delta\check{X}^{0\prime}_{t-1}) &= -\frac{\lambda^{lag}_T}{2\sqrt{T}}[\frac{B_1}{||B_1||_F||vec(\hat{B}_1)||^{\gamma}_{\infty}}, \\ &\quad \dots, \frac{B_p}{||B_p||_F||vec(\hat{B}_p)||^{\gamma}_{\infty}}] \qquad (3.55) \\ ||\sum_{t=1}^{T} -\frac{1}{\sqrt{T}} w_t \Delta\check{X}^{k\prime}_{t-1} + \frac{1}{T}(\sqrt{T}\delta^0_B)(\Delta\check{X}^0_{t-1}\Delta\check{X}^{k\prime}_{t-1})||_F &< \frac{\lambda^{lag}_T}{2\sqrt{T}}||vec(\hat{B}_k)||^{-\gamma}_{\infty} \qquad (3.56) \end{aligned}$$

for $k = p+1, \dots, P$. $\Delta\check{X}^k_{t-1}$ denotes the k-th m-dimensional block in $\Delta\check{X}_{t-1}$ and $\Delta\check{X}^0_{t-1}$ for the first p blocks. In the same way as for rank selection, define

$$\begin{aligned} V_{B_0} &= [\frac{B_1}{||B_1||_F||vec(\hat{B}_1)||^{\gamma}_{\infty}}, \dots, \frac{B_p}{||B_p||_F||vec(\hat{B}_p)||^{\gamma}_{\infty}}] \\ S_{wx0} &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T} w_t \Delta\check{X}^{0\prime}_{t-1} \qquad S_{x0} = \frac{1}{T}\sum_{t=1}^{T}\Delta\check{X}^0_{t-1}\Delta\check{X}^{0\prime}_{t-1} \\ S_{wx1} &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T} w_t \Delta\check{X}^{k\prime}_{t-1} \qquad S_{x0x1} = \frac{1}{T}\sum_{t=1}^{T}\Delta\check{X}^0_{t-1}\Delta\check{X}^{k\prime}_{t-1} \end{aligned}$$

Then from the first KKT condition (3.55) we get

$$\sqrt{T}\delta^0_B = S_{wx0}S^{-1}_{x0} - \frac{\lambda^{lag}_T}{2\sqrt{T}} V_{B_0} S^{-1}_{x0} \ . \qquad (3.57)$$

With Theorem 3.3.1 for V_{B_0} and Lemma 3.A.3 for $||S^{-1}_{x0}||_2 = O_P(1)$ it holds that

$$\frac{\lambda^{lag}_T}{\sqrt{T}}||V_{B_0}S^{-1}_{x0}||_F \leqslant \frac{\lambda^{lag}_T}{\sqrt{T}}||V_{B_0}||_F||S^{-1}_{x0}||_2 = O_p(\frac{\lambda^{lag}_T}{\sqrt{T}})$$

where the assumption on λ^{lag}_T, implies that $O_p(\frac{\lambda^{lag}_T}{\sqrt{T}}) = o_P(1)$. Thus on the true active set in the lags, the effect of the penalty vanishes.

From the second part of the KKT conditions (3.56), we obtain by plugging in (3.57) the following sufficient condition for exclusion of lags larger than p, i.e. of lags which are not in the true active set

$$||vec(-S_{wx1} + S_{wx0}S^{-1}_{x0}S_{x0x1})_j||_2 < \frac{\lambda^{lag}_T}{2\sqrt{T}}\left(||vec(\hat{B}_{p+j})||^{-\gamma}_{\infty} - ||(V_{B_0}S^{-1}_{x0}S_{x0x1})_j||_F\right) \qquad (3.58)$$

where the subscript j of a vector denotes the j-th m^2 block. With Theorem 3.3.1

for V_{B_0} together with Lemma 3.A.3 for stationary $||S_{x0x1}||_2 = O_P(1)$ we find

$$||(V_{B_0} S_{x0}^{-1} S_{x0x1})_j||_F \leqslant ||V_{B_0} S_{x0}^{-1} S_{x0x1}||_F \leqslant ||V_{B_0}||_F ||S_{x0}^{-1}||_2 ||S_{x0x1}||_2 = O_p(1)$$

and similarly on the LHS that $||(S_{wx0} S_{x0}^{-1} S_{x0x1})_j||_F = O_P(1/\sqrt{T})$. So both terms are negligible in (3.58). We use that by Theorem 3.3.1 it holds that $||vec(\hat{B}_{p+j})||_\infty^{-\gamma} = O_p\Big((\sqrt{\frac{\log m}{T}})^{-\gamma}\Big)$. Then by (3.58) and setting $N_k = vec(-S_{wx1})_k$, this yields

$$\begin{aligned}
\mathbb{P}(\sqrt{\sum_{k=1}^{m^2} N_k^2} > \frac{\lambda_T^{lag}}{2\sqrt{T}}(\frac{\sqrt{\log m}}{\sqrt{T}})^{-\gamma}) &= \mathbb{P}(\sum_{k=1}^{m^2} N_k^2 > (\frac{\lambda_T^{lag}}{2\sqrt{T}}(\frac{\sqrt{\log m}}{\sqrt{T}})^{-\gamma})^2) \\
&\leqslant \sum_{k=1}^{m^2} \mathbb{P}(|N_k| > \frac{\lambda_T^{lag}}{2m\sqrt{T}}(\frac{\sqrt{\log m}}{\sqrt{T}})^{-\gamma}) \\
&\leqslant \Big(\frac{m^2 (\log m)^{\gamma/2} C_1}{\lambda_T^{lag} T^{1/2(\gamma-1)}}\Big)^2 \text{ for } 0 < C_1 < \infty
\end{aligned}$$

where Chebyshev's inequality was applied in the last line. For the second moment bound C_1, we use that due to Assumption 3.2.1. The bound then follows from Lemma 3.A.3. Hence for P fixed and due to $\frac{\lambda_T^{lag} T^{1/2(\gamma-1)}}{m^2 (\log m)^{\gamma/2}} \to \infty$, the last line implies that with probability tending to one, irrelevant lags are excluded by the proposed Lasso procedure. □

Proof of Theorem 3.4.1

Proof. We first show that $\mathbf{E}(|u_t^k|^{4+\delta})$ is bounded for all $k = 1, \ldots, m$.

Define $\widetilde{A}_l = A_l \Sigma_w^{1/2}$ with A_l from Assumption 3.4.1. Then $\sum_{l=1}^{\infty} j||\widetilde{A}_l||_F < \infty$. Denote $\tilde{a}_{l,kj}$ as the k, j-th element in $\widetilde{A}_l$. Not that the assumption $\sum_{j=1}^{\infty} j||\tilde{A}_j||_F < \infty$ implies that $\sum_{j=1}^{\infty} ||\tilde{A}_j||_2 < \infty$. Thus for every $\varepsilon > 0$ close enough to zero, there exists an N such that for all $n > N$, $||\tilde{A}_n||_2 < \varepsilon$. Therefore, for all $1 < \zeta < \infty$, we have

$$\sum_{j=1}^{N} ||\tilde{A}_j||_2^\zeta + \sum_{j=N+1}^{\infty} ||\tilde{A}_j||_2^\zeta < \sum_{j=1}^{N} ||\tilde{A}_j||_2^\zeta + \sum_{j=N+1}^{\infty} \varepsilon^\zeta < \infty$$

We use this to bound the $4 + \delta$-th moment of u_t^k for $k = 1, \ldots, m$ split up as follows

$$u_t^k = \sum_{j=1}^{m} \tilde{a}_{0,kj} e_{t,j} + \sum_{l=1}^{\infty} \sum_{j=1}^{m} \tilde{a}_{l,kj} e_{t-l,j}$$

Define the sequence $X_l = \sum_{j=1}^{m} \tilde{a}_{l,kj} e_{t-l,j}$, then applying Rosenthal's inequality yields for the fourth moment of X_l

$$\begin{aligned}E(|X_l|^{4+\delta}) &\leqslant C_X\Big(\sum_{j=1}^{m}|\tilde{a}_{l,kj}|^{4+\delta}E(|e_{t-l,j}|^{4+\delta})+(\sum_{j=1}^{m}\tilde{a}_{l,kj}^2E(e_{t-l,j}^2))^{2+\delta/2}\Big)\\ &= O_{a.s.}\Big(\sum_{j=1}^{m}|\tilde{a}_{l,kj}|^{4+\delta}+(\sum_{j=1}^{m}\tilde{a}_{l,kj}^2)^{2+\delta/2}\Big)\\ &= O_{a.s.}\Big(||\tilde{A}_l||_2^{4+\delta}+||\tilde{A}_l||_2^{4+\delta}\Big)=O_{a.s.}(1).\end{aligned}$$

With this, we get for the partial sum $\sum_{l=0}^{L}X_l$ that

$$\begin{aligned}E(|\sum_{l=0}^{L}X_l|^{4+\delta}) &\leqslant C_X^L\Big(\sum_{l=0}^{L}\sum_{j=1}^{m}|\tilde{a}_{l,kj}|^{4+\delta}E(|e_{t-l,j}|^{4+\delta})+(\sum_{l=0}^{L}\sum_{j=1}^{m}\tilde{a}_{l,kj}^2E(e_{t-l,j}^2))^{2+\delta/2}\Big)\\ &= O_{a.s.}\Big(\sum_{l=0}^{L}\sum_{j=1}^{m}|\tilde{a}_{l,kj}|^{4+\delta}+(\sum_{l=0}^{L}\sum_{j=1}^{m}\tilde{a}_{l,kj}^2)^{2+\delta/2}\Big)\\ &= O_{a.s.}\Big(\sum_{l=0}^{L}||\tilde{A}_l||_2^{4+\delta}+(\sum_{l=0}^{L}||\tilde{A}_l||_2^2)^{2+\delta/2}\Big)=O_{a.s.}(1)\end{aligned}$$

For the $L_{4+\delta}$-convergence of $\sum_{l=0}^{L}X_l$, only remains to show that the partial sum $\sum_{l=0}^{L}X_l$ is an $L_{4+\delta}$-Cauchy sequence. Define $\xi_j=\sum_{l=0}^{j}X_l$, then for $i<j$, as i goes to infinity,

$$\begin{aligned}E(|\xi_i-\xi_j|^{4+\delta}) &= E(|\sum_{l=i+1}^{j}X_l|^{4+\delta})\\ &\leqslant C_\xi\Big(\sum_{l=i+1}^{j}E|X_l|^{4+\delta}+(\sum_{l=i+1}^{j}E(X_l^2))^{2+\delta/2}\Big)\\ &= O_{a.s.}\Big(\sum_{l=i+1}^{\infty}||\tilde{A}_j||_2^{4+\delta}+(\sum_{l=i+1}^{\infty}||\tilde{A}_l||_2^2)^{2+\delta/2}\Big)=o_{a.s.}(1)\end{aligned}$$

Therefore, ξ_j constitutes an $L_{4+\delta}$-Cauchy sequence and thus ξ_j is $L_{4+\delta}$ convergent. Therefore with dominated convergence,

$$\begin{aligned}E(|u_t^k|^{4+\delta}) &= \lim_{L\to\infty}E(|\sum_{l=0}^{L}X_l+\sum_{l=L+1}^{\infty}X_l|^{4+\delta})\\ &\leqslant \lim_{L\to\infty}C\left(E(|\sum_{l=0}^{L}X_l|^{4+\delta})+E(|\sum_{l=L+1}^{\infty}X_l|^{4+\delta})\right)<\infty\end{aligned}$$

which is the first claim of the theorem.

If the iid innovation e_0 in w_t is replaced by an i.i.d. copy $\dot{e}_0$, its impact at time t is $A_t(w_0 - \dot{w}_0) = A_t\Sigma_w^{1/2}(e_0 - \dot{e}_0) = \widetilde{A}_t(e_0 - \dot{e}_0)$. Denote the kth row of $\widetilde{A}_t$ by $\widetilde{a}_{tk}$, then it holds for $\hat{e}_0 = e_0 - e_0'$ that

$$\begin{aligned}\left(E(|\widetilde{a}_{tk}\hat{e}_0|^4)\right)^{1/4} &\leqslant \left(C_4 E(\sum_{j=1}^{m} |\widetilde{a}_{tk,j}\hat{e}_{0j}|^2)^{4/2}\right)^{1/4} \\ &\leqslant C(\sum_{j=1}^{m} \widetilde{a}_{tk,j}^2)^{1/2} = ||A_t\Sigma_w^{1/2}||_F \leqslant ||A_t||_F||\Sigma_w^{1/2}||_2 \qquad (3.59)\end{aligned}$$

by Marcinkiewicz-Zygmund inequality since each element in e_t has bounded 4-th moment. Then according to Subsection 3.1 in Wu (2007), (3.59) bounds the physical dependence measure γ_{tk} of Chen et al. (2013) elementwise. Thus we get for each element k by Assumption 3.4.1 that

$$\sum_{t=0}^{\infty} t\gamma_{tk} \leqslant ||\Sigma_w^{1/2}||_2 \sum_{t=0}^{\infty} t||A_t||_F < \infty$$

which is the sufficient condition for the elementwise strong invariance principle in Corollary 4 in Wu (2007). This implies the claim of the theorem. Moreover, the covariance matrix of $\mathbf{M}(s)$ is obtained as $\sum_{j=0}^{\infty} A_j\Sigma_w A_j'$ by elementary calculations.

□

Proof for Theorem 3.4.2

Proof. To derive the results in Theorem 3.4.2, we first show the following block-wise convergence results as in Theorem 3.2.1:

$$\begin{aligned}&||\frac{1}{T}\sum_{t=1}^{T}\Delta Z_t Z_{t-1}' - \begin{bmatrix} (\beta'\alpha)\Sigma_{z1} + \Gamma_{v1z1}^{1} & -(\beta'\alpha + I_r)\Gamma_{v2z1}^{1\prime} - \Sigma_{v1v2} \\ \Gamma_{v2z1}^{1} & \int_0^1 d\mathbf{M}_2(s)\mathbf{M}_2(s)' + \Gamma_{22}^{0}\end{bmatrix}||_F = O_P(a_n) \\ &||D_T^{-1}\frac{1}{T}\sum_{t=1}^{T} Z_{t-1}Z_{t-1}' - \begin{bmatrix} \Sigma_{z1} & -(\beta'\alpha)^{-1}\big((\beta'\alpha + I_r)\Gamma_{v2z1}^{1,\prime} + \Sigma_{v1v2} + \Gamma_{12}^{0} + \int_0^1 d\mathbf{M}_1(s)\mathbf{M}_2'(s)\big) \\ 0 & \int_0^1 \mathbf{M}_2(s)\mathbf{M}_2'(s)ds\end{bmatrix}||_F \\ = \;& O_P(a_n)\end{aligned}$$

with $a_n = \sqrt{\frac{r^2}{T}} + \sqrt{\frac{mr}{T}} + \sqrt{\frac{m^2(\log T)^{3/2}(\log\log T)}{T^{1/2}}}$.

In a similar way as in the proof for Theorem 3.2.1, we proceed with the eight blocks and highlight the differences.

1.+2. purely stationary blocks $\bar{b}_{11} = \frac{1}{T}\Delta Z_1 MZ_{1,-1}'$ *and* $\bar{\chi}_{11} = \frac{1}{T}Z_{1,-1}MZ_{1,-1}'$

For the second block, it follows from Lemma 3.A.2 that

$$||\frac{1}{T}\sum_{t=1}^{T} Z_{1,t-1}Z'_{1,t-1} - \Sigma_{z1}||_F = O_p(\frac{r}{\sqrt{T}}) \ . \tag{3.60}$$

For the first term we get from (3.15)

$$\frac{1}{T}\sum_{t=1}^{T}\Delta Z_{1,t}Z'_{1,t-1} = (\beta'\alpha)\frac{1}{T}\sum_{t=1}^{T} Z_{1,t-1}Z'_{1,t-1} + \frac{1}{T}\sum_{t=1}^{T} v_{1,t}Z'_{1,t-1}$$

which implies with (3.60) that

$$||\frac{1}{T}\sum_{t=1}^{T}\Delta Z_{1,t}Z_{1,t-1} - (\beta'\alpha)\Sigma_{z1} - \Gamma^1_{v1z1}||_F = O_p(\frac{r}{\sqrt{T}}) \tag{3.61}$$

3. mixed stationary/nonstationary block $\bar{b}_{12} = \frac{1}{T}\Delta Z_1 Z'_{2,-1}$
From (3.15) we have that $Z_{2,t} = \sum_{s=1}^{t} v_{2,s}$ which yields

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T}\Delta Z_{1,t}Z'_{2,t-1} &= -\frac{1}{T}\sum_{t=1}^{T-1} Z_{1,t}\Delta Z'_{2,t} + \frac{1}{T}(Z_{1,T}Z'_{2,T} - Z_{1,0}Z'_{2,0}) \\ &= -\frac{1}{T}(\beta'\alpha + I_r)\sum_{t=1}^{T} Z_{1,t-1}v'_{2,t} - \frac{1}{T}\sum_{t=1}^{T} v_{1,t}v'_{2,t} + R_8\end{aligned}$$

with $||R_8||_F = O_P(\sqrt{mr}/\sqrt{T})$. Hence

$$||\frac{1}{T}\sum_{t=0}^{T}\Delta Z_{1,t}Z'_{2,t-1} + (\beta'\alpha + I_r)\Gamma^{1\prime}_{v2z1} + \Sigma_{v1v2}||_F = O_p(\frac{\sqrt{mr}}{\sqrt{T}}) \tag{3.62}$$

4. mixed stationary/nonstationary block $\bar{b}_{12} = \frac{1}{T}\Delta Z_2 Z'_{1,-1}$
With $Z_{2,t} = \sum_{s=1}^{t} v_{2,s}$ from (3.15) we get that $\frac{1}{T}\sum_{t=1}^{T}\Delta Z_{2,t}Z'_{1,t-1} = \frac{1}{T}\sum_{t=1}^{T} v_{2,t}Z'_{1,t-1}$ which leads to

$$||\frac{1}{T}\sum_{t=1}^{T}\Delta Z_{2,t}Z'_{1,t-1} - \Gamma^1_{v2z1}||_F = O_p(\frac{\sqrt{mr}}{\sqrt{T}}) \tag{3.63}$$

5. purely nonstationary block $\bar{b}_{22} = \frac{1}{T}\Delta Z_2 Z'_{2,-1}$
Different from the block b_{22} in the proof for Theorem 3.2.1,the increment of $Z_{2,t}$ is no longer independent of $\mathcal{F}_{t-1}$ due to the weak dependence of $v_t = Qu_t$. Therefore, the standard discrete approximation of the stochastic integral (see, e.g. Section 2.5 of Chung and Williams (1990)) can not be directly applied.

With $Z_{2,t} = \sum_{s=1}^{t} v_{2,s}$ from (3.15) we get that

$$\frac{1}{T}\sum_{t=1}^{T}\Delta Z_{2,t}Z'_{2,t-1} = \frac{1}{T}\sum_{t=1}^{T} v_{2,t}Z'_{2,t-1} = \frac{1}{T}\sum_{t=1}^{T} v_{2,t}\sum_{s=0}^{t-1} v'_{2,s}$$

Define $\Upsilon_t = \sum_{s=0}^{t} u_s$, due to the assumption that $\sum_{j=1}^{\infty} j||A_j||_F < \infty$, for some $K > 0$, it holds that

$$\begin{aligned} u_t\Upsilon'_{t-1} = w_t\Upsilon'_{t-1} &+ A_1(w_{t-1}u'_{t-1} + w_{t-1}\Upsilon'_{t-2}) \\ &+ A_2\Big(w_{t-2}(u'_{t-1} + u'_{t-2}) + w_{t-2}\Upsilon'_{t-3}\Big) \\ &+ \dots \\ &+ A_K\Big(w_{t-K}(\sum_{j=1}^{K} u'_{t-j}) + w_{t-K}\Upsilon'_{t-K-1}\Big) + o(1) \end{aligned}$$

By summing up $u_t\Upsilon'_{t-1}$ over t and dividing the sum by T, the term $\frac{1}{T}\sum_{t=1}^{T} A_k w_{t-k}(\sum_{j=1}^{k} u'_{t-j})$ (for $1 \leqslant k \leqslant K$) satisfies

$$||\frac{1}{T}\sum_{t=1}^{T} A_k w_{t-k}(\sum_{j=1}^{k} u'_{t-j}) - A_k\Sigma_w(\sum_{j=1}^{k} A'_{k-j})||_F = O_p(||A_k||_F\frac{m}{\sqrt{T}})$$

We leave the $||A_k||_F$ in the convergence rate so that the sequence still converge at the rate of $\frac{m}{\sqrt{T}}$ after summing over k.

Sum up $A_k\Sigma_w(\sum_{j=1}^{k} A'_{k-j})$ over $k = 1, \dots, K$, i.e.,

$$\begin{aligned} \sum_{k=1}^{K} A_k\Sigma_w(\sum_{j=1}^{k} A'_{k-j}) &= A_1\Sigma_w A'_0 \\ &+ A_2\Sigma_w A'_1 + A_2\Sigma_w A'_0 \\ &+ \dots \\ &+ A_K\Sigma_w A'_{K-1} + A_K\Sigma_w A'_{K-2} + \dots + A_K\Sigma_w A'_0 \\ &\to \Gamma_u(1) + \Gamma_u(2) + \dots + \Gamma_u(K) \end{aligned}$$

where $\Gamma_u(k) = \mathbf{E}(u_t u'_{t-k})$. The left terms can be summed up over t and expressed as

$$\frac{1}{T}\sum_{t=1}^{T}\Big(\sum_{j=0}^{K} A_j w_{t-j}\Upsilon'_{t-j-1}\Big) = \sum_{j=0}^{K} A_j\Big(\frac{1}{T}\sum_{t=1}^{T} w_{t-j}\Upsilon'_{t-j-1}\Big)$$

By the same argument as in proof for Theorem 3.2.1, we conclude that

$$||\frac{1}{T}\sum_{t=1}^{T} w_{t-j}\Upsilon'_{t-j-1} - \int_0^1 d\mathbf{M}_w\mathbf{M}'||_F = O_p(\frac{m(\log T)^{3/4}(\log\log T)^{1/2}}{T^{1/4}})$$

where $\mathbf{M}_w$ denotes the m-dimensional Brownian motion with the same covarianc matrix as w_t. According to Theorem 3.4.1, we can conclude that

$$||\frac{1}{T}\sum_{t=1}^{T} u_t\Upsilon'_{t-1} - \int_0^1 d\mathbf{M}\mathbf{M}' - \sum_{k=1}^{\infty}\Gamma_u(k)||_F = O_p(\frac{m(\log T)^{3/4}(\log\log T)^{1/2}}{T^{1/4}}) \quad (3.64)$$

The convergence rate of those terms to $\sum_{k=1}^{\infty}\Gamma_u(k)$ is $\frac{m}{\sqrt{T}}$, dominated by the rate of strong invariance principle and thus ignored here.

The desired result can be achieved by pre- and post-multiplying $u_t\Upsilon'_{t-1}$ by β' or $\alpha'_\perp$.

6. mixed stationary/nonstationary block $\bar{\chi}_{12} = \frac{1}{T}Z_{1,-1}Z'_{2,-1}$
From (3.15) we get that

$$\frac{1}{T}\sum_{t=1}^{T}\Delta Z_{1,t}Z'_{2,t-1} = \frac{1}{T}\sum_{t=1}^{T}(\beta'\alpha)Z_{1,t-1}Z'_{2,t-1} + \frac{1}{T}\sum_{t=1}^{T}v_{1,t}Z'_{2,t-1}$$

which we rearrange as

$$\frac{1}{T}\sum_{t=1}^{T}Z_{1,t-1}Z'_{2,t-1} = (\beta'\alpha)^{-1}(\frac{1}{T}\sum_{t=1}^{T}\Delta Z_{1,t}Z'_{2,t-1} - \frac{1}{T}\sum_{t=1}^{T}v_{1,t}Z'_{2,t-1}) .$$

Using (3.62) for the first term on the right, the strong invariance principle of Theorem 3.4.1 for the second term as above and $||(\beta'\alpha)^{-1}||_2 = O(r^{\tau_1})$ by Assumption 3.4.2 it holds that

$$\begin{aligned}&||\frac{1}{T}\sum_{t=1}^{T}Z_{1,t-1}Z'_{2,t-1} + (\beta'\alpha)^{-1}\Big((\beta'\alpha + I_r)\Gamma^{1\prime}_{v2z1} + \Sigma_{v1v2} + \Gamma^0_{12} + \int_0^1 d\mathbf{M}_1\mathbf{M}'_2\Big)||_F \\ = \quad & O\Big(r^{\tau_1}\sqrt{\frac{mr}{T} + \frac{mr(\log T)^{3/2}(\log\log T)}{\sqrt{T}}}\Big) \quad (3.65)\end{aligned}$$

7. mixed stationary/nonstationary block $\bar{\chi}_{21} = \frac{1}{T}\left(\frac{1}{T}Z_{2,-1}Z'_{1,-1}\right)$
By similar argument as in the independent case for block χ_{21}, it is sufficient to work here withe conservative upper bound $O_P(r^{\tau_1})$ from (3.65) for each element in the

inner bracket. We thus obtain

$$||\frac{1}{T^2}\sum_{t=1}^{T} Z_{2,t-1}Z'_{1,t-1}||_F = O_p(\frac{\sqrt{mr}r^{\tau_1}}{T}) \tag{3.66}$$

8. purely nonstationary block $\bar{\chi}_{22} = \frac{1}{T}\left(\frac{1}{T}Z_{2,-1}Z'_{2,-1}\right)$
As before, we show the distance between $\frac{1}{T^2}\sum_{t=1}^{T} Z_{2,t-1}Z'_{2,t-1}$ and $\int_0^1 d\mathbf{M}_2\mathbf{M}_2$. Element-wise, we have

$$\begin{aligned}
& \frac{1}{T^2}\sum_{t=1}^{T} Z^i_{2,t-1}Z^j_{2,t-1} - \int_0^1 \mathbf{M}_{2,i}(s)\mathbf{M}_{2,j}(s)ds \\
= & \sum_{t=1}^{T}\int_{\frac{t-1}{T}}^{\frac{t}{T}} (\frac{1}{\sqrt{T}}Z^i_{2,t-1})\Big(\frac{1}{\sqrt{T}}Z^j_{2,t-1} - \mathbf{M}_{2,j}(s)\Big)ds \\
+ & \sum_{t=1}^{T}\int_{\frac{t-1}{T}}^{\frac{t}{T}} \Big(\frac{1}{\sqrt{T}}Z^i_{2,t-1} - \mathbf{M}_{2,i}(s)\Big)(\mathbf{M}_{2,j}(s))ds
\end{aligned}$$

We have shown that $|\frac{1}{\sqrt{T}}Z^i_{2,t-1} - \mathbf{M}_{2,i}(\frac{t-1}{T})| = O_p(\frac{(\log T)^{3/4}(\log\log T)^{1/2}}{T^{1/4}})$ and for any Brownian motion element in $\mathbf{M}(s)$, $\max_{\frac{t-1}{T}\leqslant s\leqslant\frac{t}{T}} |\mathbf{M}_j(s) - \mathbf{M}_j(\frac{t-1}{T})| = O_p(\sqrt{\frac{\log T}{T}})$. Thus in total, we get

$$||\frac{1}{T^2}Z_{2,t-1}Z'_{2,t-1} - \int_0^1 \mathbf{M}_2\mathbf{M}'_2||_F \quad = \quad O_p(m\frac{(\log T)^{3/4}(\log\log T)^{1/2}}{T^{1/4}}) \tag{3.67}$$

Now the first part of the initial claim follows from (3.61)-(3.64) and the second part from (3.60) and (3.65)-(3.67). In the same manner as the proof of Theorem 3.2.1, we can define $\bar{\chi}$ composed of the blocks $\bar{\chi}_{11} - \bar{\chi}_{22}$ and $\xi = \bar{\chi}^{-1}$. Then the final result for $\widetilde{\Psi}$ follows from direct calculations. □

Proof of Theorem 3.4.3

Proof. Define $\beta_0 = \begin{bmatrix} \beta' \\ \beta'_\perp \end{bmatrix}$. We thus obtain for $\beta_0\widetilde{\Pi}'$

$$\begin{aligned}
\begin{pmatrix} \beta'\widetilde{\Pi}' \\ \beta'_\perp\widetilde{\Pi}' \end{pmatrix} &= \begin{pmatrix} I_r & \frac{1}{T}\beta'\alpha_\perp \\ 0 & \frac{1}{T}\beta'_\perp\alpha_\perp \end{pmatrix}\Big(Q^{-1}\widetilde{\Psi}\Big)' \qquad (3.68) \\
&= \begin{pmatrix} I_r & \frac{1}{T}\beta'\alpha_\perp \\ 0 & \frac{1}{T}\beta'_\perp\alpha_\perp \end{pmatrix}\begin{pmatrix} \alpha(\beta'\alpha)^{-1}\widetilde{\Psi}_{11} + \beta_\perp(\alpha'_\perp\beta_\perp)^{-1}\widetilde{\Psi}_{21} & \alpha(\beta'\alpha)^{-1}\widetilde{\Psi}_{12} + \beta_\perp(\alpha'_\perp\beta_\perp)^{-1}\widetilde{\Psi}_{22} \end{pmatrix}'
\end{aligned}$$

From (3.68) and Theorem 3.4.2, we get

$$\begin{aligned} ||\beta'\widetilde{\Pi}' - \alpha'_\star||_F &= O_p(\sqrt{\frac{mr}{T}}) && (3.69) \\ ||\beta'_\perp\widetilde{\Pi}'||_2 &= O_p(\frac{r^{\tau_1}}{T}) && (3.70) \end{aligned}$$

The l_2 norms of $\widetilde{\Psi}_{21}$ and $\widetilde{\Psi}_{22}$ may increase with r^{τ_1}, which slows down the converging rate of the irrelevant basis.

Due to the unitary invariant property of singular values, we have

$$\sigma_j(\beta_0\widetilde{\Pi}') = \sigma_j(S\widetilde{\Pi}') = \sigma_j(\widetilde{R})$$

which implies that

$$|\sigma_j(\widetilde{R}) - \sigma_j(\alpha_\star)| = O_p(\sqrt{\frac{mr}{T}}) \quad \text{for } j = 1, \ldots, r \tag{3.71}$$

by matrix perturbation theory (Mirsky version, Theorem 4.11 of Stewart and Sun (1990)) and (3.69). The column-pivoting step in QR decomposition makes the $\widetilde{R}_{11}$ a well-conditioned matrix, thus the largest r singular values in $\widetilde{R}$ are contributed by $\widetilde{R}_1$, the first r-columns. Besides, the upper-triangular structure of $\widetilde{R}$ excludes linear dependence between any two rows. Therefore, we can conclude that

$$\sigma_r(\widetilde{R}) \leqslant \sqrt{\sum_{j=k}^{m} \widetilde{R}(k,j)^2} \leqslant \sigma_1(\widetilde{R}) \quad \text{for } k = 1, \ldots, r \tag{3.72}$$

The matrix perturbation theory result (3.71) provides further bounds for l_2 norm of each row in $\widetilde{R}_1$, i.e.,

$$\begin{aligned} \sigma_r(\widetilde{R}) &\geqslant \sigma_r(\alpha_\star) - O_p(\sqrt{\frac{mr}{T}}) \\ \sigma_1(\widetilde{R}) &\leqslant \sigma_1(\alpha_\star) + O_p(\sqrt{\frac{mr}{T}}) \end{aligned}$$

Also by the upper-triangular structure and column-pivoting, we can derive that

$$\sqrt{\sum_{j=k}^{m} \widetilde{R}(k,j)^2} = O_p(\frac{r^{\tau_1}}{T}) \quad \text{for } k = r+1, \ldots, m \tag{3.73}$$

Moreover, (3.73) leads to the conclusion that

$$||\widetilde{R}_{22}||_F = O_p(\frac{\sqrt{m}}{T}r^{\tau_1}) \tag{3.74}$$

The difference between $\widetilde{R}_1$ and $\widetilde{R}$ is $\widetilde{R}_{22}$. Therefore, we can also conclude

$$|\sigma_j(\widetilde{R}_1) - \sigma_j(\widetilde{R})| = O_p(\frac{\sqrt{m}r^{\tau_1}}{T}), \qquad j = 1, .., r$$

and thus

$$|\sigma_j(\widetilde{R}_1) - \sigma_j(\alpha_\star)| = O_p(\sqrt{\frac{mr}{T}}) \quad \text{for } j = 1, \dots, r \tag{3.75}$$

(3.68) can be further written as

$$\begin{pmatrix} \beta'\widetilde{\Pi}' \\ \beta'_\perp\widetilde{\Pi}' \end{pmatrix} = \begin{pmatrix} \beta'\widetilde{S}_1\widetilde{R}_{11} & \beta'\widetilde{S}_1\widetilde{R}_{12} + \beta'\widetilde{S}_2\widetilde{R}_{22} \\ \beta'_\perp\widetilde{S}_1\widetilde{R}_{11} & \beta'_\perp\widetilde{S}_1\widetilde{R}_{12} + \beta'_\perp\widetilde{S}_2\widetilde{R}_{22} \end{pmatrix} \tag{3.76}$$

with the after QR-decomposition components.

By equating the (3.68) and (3.76) we also have

$$\begin{pmatrix} \beta'_\perp\widetilde{S}_1\widetilde{R}_{11} & \beta'_\perp\widetilde{S}_1\widetilde{R}_{12} + \beta'_\perp\widetilde{S}_2\widetilde{R}_{22} \end{pmatrix} = \frac{1}{T}(\beta'_\perp\alpha_\perp)\Big(\alpha(\beta'\alpha)^{-1}\widetilde{\Psi}_{12} + \beta_\perp(\alpha'_\perp\beta_\perp)^{-1}\widetilde{\Psi}_{22}\Big)'$$

which is equivalent to

$$\begin{aligned} \beta'_\perp\widetilde{S}_1 &= -\Big[\ 0 \quad \beta'_\perp\widetilde{S}_2\widetilde{R}_{22}\ \Big]\widetilde{R}'_1(\widetilde{R}_1\widetilde{R}'_1)^{-1} \\ &+ \frac{1}{T}(\beta'_\perp\alpha_\perp)\Big(\alpha(\beta'\alpha)^{-1}\widetilde{\Psi}_{12} + \beta_\perp(\alpha'_\perp\beta_\perp)^{-1}\widetilde{\Psi}_{22}\Big)'\widetilde{R}'_1(\widetilde{R}_1\widetilde{R}'_1)^{-1} \end{aligned}$$

The singular values of $\widetilde{R}_1$ can be approximated by those of α. Therefore we conclude that $||\beta'_\perp\widetilde{S}_1||_F = O_p(\frac{\sqrt{m}r^{\tau_1+2\tau_2}}{T})$.

□

Proof of Theorem 3.4.4

Proof. Denote $\widetilde{S}'Y_{t-1} = \begin{bmatrix} \breve{Z}_{1,t-1} \\ \breve{Z}_{2,t-1} \end{bmatrix}$ where $\breve{Z}_{1,t-1}$ is the projection of Y_{t-1} onto the subspace generated by $\widetilde{S}_1$. Because the distance between $\widetilde{S}_1$ and β converges at the rate of T, faster than other error terms mentioned above, we use $Z_{1,t-1}$ instead of $\breve{Z}_{1,t-1}$ in this proof. While both $\breve{Z}_{2,t-1}$ and $Z_{2,t-1}$ are non-stationary process, we

can also use $Z_{2,t-1}$ instead of $\breve{Z}_{2,t-1}$ to keep the proof easier to read. Besides, we do not distinguish different matrix representations of $\hat{\alpha}$.
Define $\alpha_0 = [\alpha, 0], \hat{\alpha} = \alpha + \Sigma_{uz1}\Sigma_{z1}^{-1}, \hat{\alpha}_0 = [\hat{\alpha}, 0_{m\times m-r}], \quad \hat{u}_t = u_t - (\hat{\alpha}-\alpha)Z_{1,t-1}$. Then $\mathbf{E}(Z_{1,t-1}\hat{u}_t') = 0$. δ_R and δ_{R1} are defined as before. We have the same Lasso criterion function as in Theorem 3.2.3 which leads to the identical KKT optimality conditions Thus the Karush-Kuhn-Tucker (KKT) condition for group-wise variable selection from (3.45) is

$$-\frac{1}{\sqrt{T}}\sum_{t=1}^{T} w_t \widetilde{Z}_{1,t-1}' + \sqrt{T}\delta_{R1}\frac{1}{T}\sum_{t=1}^{T}\widetilde{Z}_{1,t-1}\widetilde{Z}_{1,t-1}' = -[\frac{\bar{\lambda}_{1,T}}{2\sqrt{T}}\frac{\bar{\alpha}(,1)}{||\bar{\alpha}(,1)||_2}, \dots, \frac{\bar{\lambda}_{r,T}}{2\sqrt{T}}\frac{\bar{\alpha}(,r)}{||\bar{\alpha}(,r)||_2} \quad (3.77)$$

$$||\Big(\sum_{t=1}^{T} -2\frac{1}{T}w_t\widetilde{Z}_{2,t-1}' + 2\sqrt{T}\delta_{R1}\frac{1}{T^{3/2}}\widetilde{Z}_{1,t-1}\widetilde{Z}_{2,t-1}'\Big)_j||_2 < \frac{\bar{\lambda}_{r+j,T}}{T} \quad (3.78)$$

where $\bar{\lambda}_{j,T} = \frac{\lambda_T^{rank}}{\tilde{\mu}_j^{\gamma}}$ and the subscript j denotes the jth column.
According to the definition of $\hat{u}_t$, $||\frac{1}{T}\sum_{t=1}^{T}\hat{u}_t Z_{1,t-1}'||_F = O_p(\sqrt{\frac{mr}{T}})$.

Rewrite

$$[\frac{\bar{\lambda}_{1,T}}{2\sqrt{T}}\frac{\hat{\alpha}(,1)}{||\hat{\alpha}(,1)||_2}, \dots, \frac{\bar{\lambda}_{r,T}}{2\sqrt{T}}\frac{\hat{\alpha}(,r)}{||\hat{\alpha}(,r)||_2}] = \frac{\lambda_T^{rank}}{2\sqrt{T}}[\frac{\hat{\alpha}(,1)}{||\hat{\alpha}(,1)||_2\tilde{\mu}_1^{\gamma}}, \dots, \frac{\hat{\alpha}(,r)}{||\hat{\alpha}(,r)||_2\tilde{\mu}_r^{\gamma}}] = \frac{\lambda_T^{rank}}{2\sqrt{T}}V_\alpha$$

Define

$$S_{z1z1} = \frac{1}{T}\sum_{t=1}^{T} Z_{1,t-1}Z_{1,t-1}' \qquad S_{uz1} = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\hat{u}_t Z_{1,t-1}'$$

$$S_{z1z2} = \frac{1}{T^{3/2}}\sum_{t=1}^{T} Z_{1,t-1}Z_{2,t-1}' \qquad S_{uz2} = \frac{1}{T}\sum_{t=1}^{T}\hat{u}_t Z_{2,t-1}'$$

Then we can derive that

$$\sqrt{T}\delta_{R1} = -\frac{\lambda_T^{rank}}{2\sqrt{T}}V_\alpha S_{z1z1}^{-1} + S_{uz1}S_{z1z1}^{-1}$$

as $\frac{\lambda_T^{rank} r^{\tau_2\gamma+\frac{1}{2}}}{\sqrt{T}} \to 0$, $||\delta_{R1}||_F = O_p(\sqrt{\frac{mr}{T}})$.

To study the tail properties of elements in S_{uz2}, we need the following results based on Beveridge-Nelson decomposition.

$$\begin{aligned}
u_{t-q}u_t' &= (\sum_{j=0}^{\infty} A_j w_{t-q-j})(\sum_{k=0}^{\infty} A_k w_{t-k})' = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} A_j w_{t-q-j} w_{t-k}' A_k' \\
&= \sum_{j=q}^{\infty}\sum_{k=0}^{q-1} A_{j-q} w_{t-j} w_{t-k}' A_k' + \sum_{j=q}^{\infty}\sum_{k=q}^{\infty} A_{j-q} w_{t-j} w_{t-k}' A_k' \\
&= \sum_{j=q}^{\infty}\sum_{k=0}^{q-1} A_{j-q} w_{t-j} w_{t-k}' A_k' + \sum_{k=q}^{\infty} A_{k-q} w_{t-k} w_{t-k}' A_k' \\
&+ \sum_{j=q}^{\infty}\sum_{i=1}^{\infty} A_{j+i-q} w_{t-j-i} w_{t-j}' A_j' + \sum_{j=q}^{\infty}\sum_{i=1}^{\infty} A_{j-q} w_{t-j} w_{t-j-i}' A_{j+i}'
\end{aligned}$$

The term $\sum_{k=q}^{\infty} A_{k-q} w_{t-k} w_{t-k}' A_k'$ can be further decomposed as

$$\begin{aligned}
& \sum_{k=q}^{\infty} A_{k-q} w_{t-k} w_{t-k}' A_k' \\
= & \sum_{k=q}^{\infty} A_{k-q} w_{t-q} w_{t-q}' A_k' - \sum_{k=q+1}^{\infty} A_{k-q} w_{t-q} w_{t-q}' A_k' \\
+ & \sum_{k=q+1}^{\infty} A_{k-q} w_{t-q-1} w_{t-q-1}' A_k' - \sum_{k=q+2}^{\infty} A_{k-q} w_{t-q-1} w_{t-q-1}' A_k' \\
+ & \ldots \\
+ & \sum_{k=q+K}^{\infty} A_{k-q} w_{t-q-K} w_{t-q-K}' A_k' - \sum_{k=q+K+1}^{\infty} A_{k-q} w_{t-q-K} w_{t-q-K}' A_k' \\
+ & \ldots \\
= & \sum_{k=q}^{\infty} A_{k-q} w_{t-q} w_{t-q}' A_k' \\
- & (\sum_{k=q+1}^{\infty} A_{k-q} w_{t-q} w_{t-q}' A_k' - \sum_{k=q+1}^{\infty} A_{k-q} w_{t-q-1} w_{t-q-1}' A_k') \\
- & \ldots \\
- & (\sum_{k=q+K+1}^{\infty} A_{k-q} w_{t-q-K} w_{t-q-K}' A_k' - \sum_{k=q+K+1}^{\infty} A_{k-q} w_{t-q-K-1} w_{t-q-K-1}' A_k') \\
- & \ldots
\end{aligned}$$

Therefore, if we sum up $\sum_{k=q}^{\infty} A_{k-q} w_{t-k} w_{t-k}' A_k'$ over t, only $\sum_{k=q}^{\infty} A_{k-q} w_{t-q} w_{t-q}' A_k'$

remains and the other terms get deleted.

$$\begin{aligned} & \frac{1}{T}\sum_{t=1}^{T}\sum_{k=q}^{\infty} A_{k-q} w_{t-k} w'_{t-k} A'_k \\ = & \frac{1}{T}\sum_{t=1}^{T}\sum_{k=q}^{\infty} A_{k-q} w_{t-q} w'_{t-q} A'_k + O_p(\frac{1}{T}) \\ = & \sum_{k=q}^{\infty} A_{k-q}(\frac{1}{T}\sum_{t=1}^{T} w_{t-q} w'_{t-q}) A'_k + O_p(\frac{1}{T}) \end{aligned}$$

The expectation of $\frac{1}{T}\sum_{t=1}^{T}\sum_{k=q}^{\infty} A_{k-q} w_{t-k} w'_{t-k} A'_k$ is thus $\sum_{k=q}^{\infty} A_{k-q}\Sigma_w A'_k$. Each element in $\sum_{t=1}^{T}(w_t w'_t - \Sigma_w)$ constitute a martingale process due to the independence of w_t. The bounded second moment of elements in $w_t w'_t - \Sigma_w$ is ensured by the $(4+\delta)$-th moment condition for w_t or e_t. The martingale property of the other terms in $\sum_{t=1}^{T} u_{t-q} u'_t$ can be proved similarly since given $i \neq j$, $\frac{1}{T}\sum_{t=1}^{T} w_{t-i} w'_{t-j}$ converges to zero due to the independent w_t. To sum up, we have shown that each element in

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Big(u_{t-q}u'_t - \mathbf{E}(u_{t-q}u'_t)\Big) \tag{3.79}$$

has bounded second moment. Because this result holds for a general $q \geqslant 0$, after linear combination, we can also show that each element in

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Big(Z_{1,t-1}u'_t - \mathbf{E}(Z_{1,t-1}u'_t)\Big)$$

has bounded second moment. This is also true if we sum up (3.79) over q. However, to ensure the convergence, we must divide the new result by $\sqrt{T}$, i.e., each element in

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Big(\frac{1}{\sqrt{T}}\sum_{q=1}^{t} u_{t-q}u'_t - \frac{1}{\sqrt{T}}\sum_{q=1}^{t}\mathbf{E}(u_{t-q}u'_t)\Big) \tag{3.80}$$

has bounded second moment. Summing up over q is well-defined due the fast convergence rate of A_j according to our assumption. Therefore, it holds also for

$$\frac{1}{T}\sum_{t=1}^{T}\Big(Z_{2,t-1}v'_t - \mathbf{E}(Z_{2,t-1}v'_t)\Big) \tag{3.81}$$

From the block 6 in the proof of Theorem 3.4.3, the l_2 norm of $\frac{1}{T}\sum_{t=1}^{T} Z_{1,t-1}Z'_{2,t-1}$ is inflated by r^{τ_1} approximating l_2 norm of $(\beta'\alpha)^{-1}$, which make it more difficult to exclude the irrelevant groups compared to the i.i.d case.

To exclude the irrelevant part, KKT condition is satisfied if

$$||(S_{uz1}S_{z1z1}^{-1}S_{z1z2} - S_{uz2})_k||_2 < \frac{\lambda_T^{rank}}{2T}\tilde{\mu}_{r+k}^{-\gamma} - \frac{\lambda_T^{rank}}{2\sqrt{T}}||(V_\alpha S_{z1z1}^{-1}S_{z1z2})_k||_2 \tag{3.82}$$

$$\begin{aligned} & \frac{\lambda_T^{rank}}{2\sqrt{T}}||(V_\alpha S_{z1z1}^{-1}S_{z1z2})_j||_2) \leqslant \frac{\lambda_T^{rank}}{2\sqrt{T}}||V_\alpha S_{z1z1}^{-1}S_{z1z2}||_F \\ \leqslant\ & \frac{\lambda_T^{rank}}{2\sqrt{T}}||V_\alpha||_F||S_{z1z1}^{-1}||_2||S_{z1z2}||_2 \\ =\ & O_p\Big(\frac{\lambda_T^{rank}r^{\tau_1+\tau_2\gamma+\frac{1}{2}}}{T}\Big) \end{aligned}$$

Thus the RHS of (3.82) is dominated by the first term. The LHS of (3.82) is dominated by S_{uz2} since the l_2 norm of $S_{uz1}S_{z1z1}^{-1}S_{z1z2}$ converges to zero at the rate of $\frac{r^{\tau_1}}{\sqrt{T}}$ as S_{z1z2}. Denoting N_i as element in S_{uz2} and $\widetilde{N}_i$ as the perturbation of $\frac{1}{T}\sum_{t=1}^{T}\hat{u}_tZ_{2,t-1}$ from the expectation. By the same argument as above, we have

$$\begin{aligned} & \mathbb{P}(\sqrt{\sum_{i=1}^{m}N_i^2} > \frac{\lambda_T^{rank}}{2T}\tilde{\mu}_{r+k}^{-\gamma}) \\ \leqslant\ & \mathbb{P}(\sum_{i=1}^{m}N_i^2 > \left(\frac{\lambda_T^{rank}}{2T}\tilde{\mu}_{r+k}^{-\gamma}\right)^2) \\ \leqslant\ & \sum_{i=1}^{m}\mathbb{P}(|N_i| > \frac{\lambda_T^{rank}}{2T\sqrt{m}}\tilde{\mu}_{r+k}^{-\gamma}) \\ \leqslant\ & \sum_{i=1}^{m}\mathbb{P}(|\widetilde{N}_i| + |c| > \frac{\lambda_T^{rank}}{2T\sqrt{m}}\tilde{\mu}_{r+k}^{-\gamma}) \\ \leqslant\ & \bar{C}_0 r^{2\tau_1} m\left(\frac{\sqrt{m}r^{\tau_1\gamma}}{\lambda_T^{rank}T^{\gamma-1}}\right)^2 = \bar{C}_0\left(\frac{mr^{\tau_1(\gamma+1)}}{\lambda_T^{rank}T^{\gamma-1}}\right)^2 \end{aligned}$$

To make all the non-stationary parts excluded from the final estimator, we require that

$$\frac{\lambda_T^{rank}T^{\gamma-1}}{m^{3/2}r^{\tau_1(\gamma+1)}} \to \infty$$

□

3.B Figures

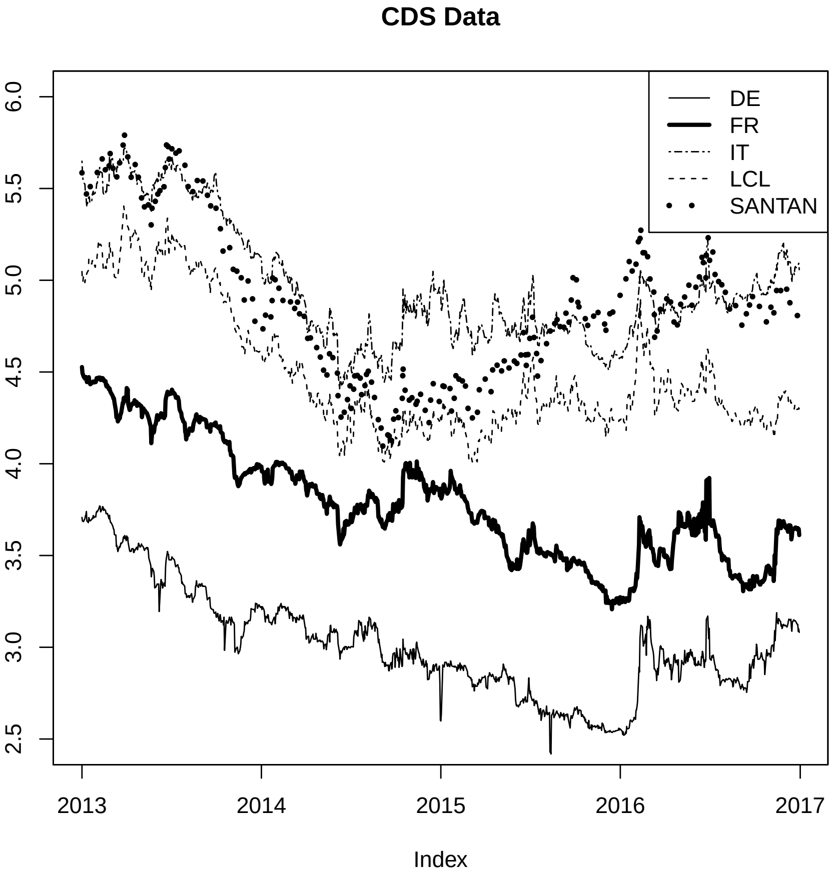


Figure 3.4. CDS data of Germany, France and Italy

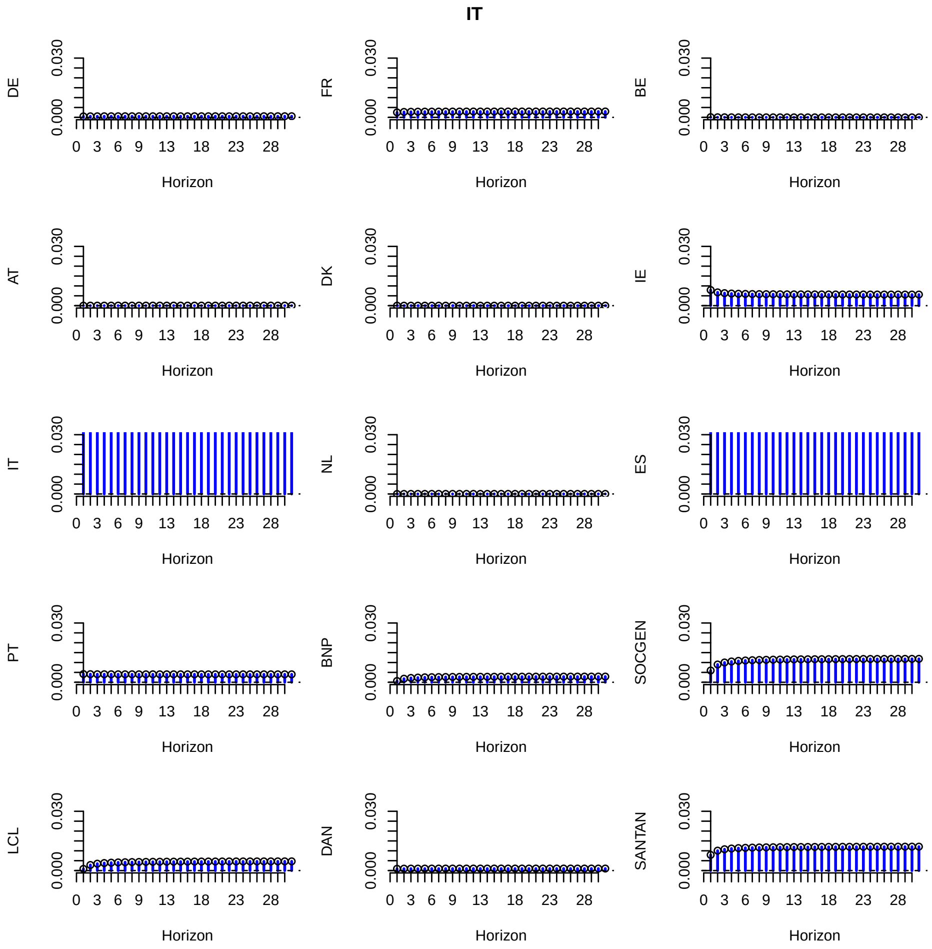


Figure 3.5. FEVD from Italy. The FEVD of Italy to itself is plotted as zero to highlight its contribution to others.

4 Pre-screening & Reduced Rank Regression in Ultra-High Dimensional Cointegration

4.1 Introduction

A new challenge for data analysis nowadays is high dimensionality, i.e., the number of variables could be large compared with the number of observations. Moreover, many data are further characterized by dynamics and non-stationarity. In theory, how to estimate a high dimensional non-stationary time series model is still an unsolved but important issue for statisticians and econometricians. Such a theoretical result has strong empirical implications, for example, to search for possible cointegration relations for statistical arbitrages from thousands of stocks.

The standard tool to model multivariate non-stationary time series is vector error correction model, as proposed by Engle and Granger (1987). The first rigorous estimation framework is Johansen's approach proposed in a sequence of papers around 1990. Although it has been the most popular approach so far, the model selection result is not consistent and the highest affordable dimension is eleven because the complicated critical values come from simulation results. In order to improve the model estimation results, several other methodologies have been proposed. One typical work is Chao and Phillips (1999), which compares different possible models by posterior information criteria. Another consistent model selection and estimation approach is proposed in Cavaliere et al. (2012) by adapting bootstrap techniques to the time series setting. In the extended version, e.g. Cavaliere et al. (2014) also shows the power of the bootstrap approach in presence of conditional heteroskedasticity. But all these methods can only be applied with small sets of variables. The new estimation strategy proposed in the previous chapter based on Lasso in high dimensions. In this way, the model selection result is consistent and detailed numerical examples (such as 20 dimensions and 400 observations) are covered in the simulation. However, their method can't be directed extended to ultra-high dimensions without further considerations. Therefore, a new method with sound statistical properties is required to estimate the high dimensional VECM.

In this paper, we determine the rank and cointegration vectors from reduced rank regression (see e.g. Anderson (2002) and RRR hence-after) when the dimension is much higher than the number of observations. For this goal, we assume the unknown coefficient matrices satisfy some sparse structures. Thus a pre-screening step will be introduced to reduce the dimension and avoid overfitting the further analysis. A pioneering paper in high dimensional statistics is Bickel et al. (2009), which derives the upper bound for the bias of Lasso estimate for additive model. Kock and Callot (2015) extends this method to high dimensional stationary vector autoregressive model. However, this approach can't be used for the estimation of cointegration vectors in non-stationary case. In order to follow the reduced rank regression approach as Anderson (2002), the ultra-high dimensionality requires pre-screening techniques as introduced in Ma (2013), which estimates the principle components in a high dimensional setting by assuming sparse factors. Gao et al. (2015) propose for the first time the rate-optimal non-asymptotic minimax estimation of the canonical correlation analysis in ultra-high dimensions. However, they fail to propose a feasible algorithm to implement their method.

In order to characterize the accuracy of the estimated cointegration vectors, which are estimated as the eigenspace for the relevant covariance matrices with large dimensions, we require the results from the matrix perturbation theory, see e.g. Stewart and Sun (1990), Li (1994) and Golub and Van Loan (2013). It should be noted that different from existing high dimensional literature on principle component analysis, high dimensional reduced rank regression proposed by Anderson (2002) requires generalized eigenvalue decomposition, which is more complicated from the perspective of matrix perturbation theory.

High dimensional time series is still an open research field with the following typical literature. A fundamental contribution is from Chen et al. (2013), which derives the convergence rate of the sample covariance matrix with non-i.i.d non-Gaussian observations based on Wu (2007), which shows the convergence rate of strong invariance principle with non-i.i.d. non-Gaussian data. Lam and Yao (2012) proposes an estimation framework based on factor model for rank selection and loading matrix estimation. Then Zhang et al. (2018) extends Lam and Yao (2012) to a cointegrated high dimensional time series and obtains similar results. But the focus of Zhang et al. (2018) is only cointegration without taking VECM structure into account. A very recent paper studying high dimensional VECM is Onatski and Wang (2018), whose analysis relies on random matrix theory and can't be extended to ultra-high dimensions. Therefore, how to estimate ultra-high dimensional VECM by eigenvalue analysis is still an open issue in literature. This paper contributes to the literature in the following respects: i) we derive the statistical properties of estimators for rank and cointegration vectors in high dimensional VECM by principal analysis; ii) we propose a feasible algorithm to pre-screen the elements in loading matrix α and cointegration matrix β with good theoretical properties.

4.2 Model

In this paper we consider the cointegrated vector autoregressive model without transient term as in (4.1),

$$\Delta Y_t = \alpha\beta' Y_{t-1} + w_t, \quad t = 1, 2, \ldots, T \tag{4.1}$$

where the error terms $\{w_t\}$ are assumed to be i.i.d. $N(0, \sigma^2 I_m)$ and $\alpha, \beta \in \mathbb{R}^{m\times r}$. Besides, we need Assumption 4.2.1 for further analysis:

Assumption 4.2.1. *We need the following assumptions for components in* (4.1)

1. *$rank(\alpha) = rank(\beta) = r$, i.e. both α and β have full-column rank.*
2. *$\beta'\beta = I_r$, i.e. β is orthonormal.*
3. *Denote the singular value fo α as $\sigma_j(\alpha)$. W.l.o.g, it is assumed that $0 < \sigma_r(\alpha) \leqslant \cdots \leqslant \sigma_2(\alpha) \leqslant \sigma_1(\alpha) < \infty$ and $\frac{1}{\sigma_r(\alpha)} < \infty$.*
4. *The eigenvalues of $I_r + \beta'\alpha$ lie within the unit circle.*
5. *The dimension m can increase with T but r is fixed.*

The last point in Assumption 4.2.1 implies that when T increases, the dimension m also explodes. In our ultra-high dimensional setting, $m >> T$ could also happen. To focus on the main challenges in the analysis, the rank r is assumed to be fixed.

In fixed dimensional case, reduced rank regression can determine (4.1) by solving the generalized eigenvalue problem

$$S_{10}S_{00}^{-1}S_{01}\hat{h}_j = \lambda_j S_{11}\hat{\lambda}_j$$

where $S_{00} = \frac{1}{T}\sum_{t=1}^{T}\Delta Y_t \Delta Y_t'$, $S_{10} = \frac{1}{T}\sum_{t=1}^{T} Y_{t-1}\Delta Y_t'$, $S_{01} = S_{10}'$, $S_{11} = \frac{1}{T}\sum_{t=1}^{T} Y_{t-1}Y_{t-1}'$. However, when the dimension is moving, the traditional analysis based on central limit theorem doesn't work here. When $m > T$, the matrix S_{00} is not even invertible. Therefore, we must have stronger assumptions in ultra-high dimensions than in Anderson (2002).

In high dimensional setting, we need to Assumptions 4.2.2 and 4.2.3 in order to impose a sparse structure on (4.1)

Assumption 4.2.2. *Each α_j, the j-th column of α, satisfies the weak l_d ball condition, i.e. $|\alpha_j|_{(k)} \leqslant c_{\alpha,j}k^{-1/d}$ for all k and $d \in (0, 2)$. The subscript (k) denotes the k-th largest absolute value in α_j.*

Assumption 4.2.3. *The cointegrating space spanned by β satisfies the following assumptions*

1. *For a bounded positive number $K < \infty$, only K rows in β have nonzero elements, i.e., $\sum_{k=1}^{m} \mathbb{I}_{\max_{j \leqslant r} |\beta(k,j)| \neq 0} = K$. Denote the set of indices for which the corresponding rows in β have nonzero elements as $\mathbf{M}_K$, i.e.*

$$\mathbf{M}_K = \{k : \max_{j \leqslant r} |\beta(k,j)| \neq 0\}$$

Without loss of generality, we further assume that $\mathbf{M}_K = \{1, 2, \cdots, K\}$ by re-ordering Y_t.

2. *There exists a $\varepsilon_\beta > 0$ significantly different from zero thus that $\max_{j \leqslant r} |\beta(i,j)| > \varepsilon_\beta$ for all $i \in \mathbf{M}_K$, $j \leqslant r$.*

It should be noted that the matrix representation of β is not unique but Assumption 4.2.3 is invariant after β being post-multiplied by an orthonormal matrix.

4.3 Pre-screening Techniques

4.3.1 Pre-screening for loading matrix

By pre-selecting the loading matrix α, we try to identify those rows with at least one element significantly different from zero. If the k-th row in α has all the elements very close to zero, then ΔY_t^k is mainly determined by the innovation w_t^k and behaves quite similar to white noise. In such a situation, we exclude these variables from reduced rank regression since they provide little valuable information about the canonical correlation.

Denote α^k as the k-th row of α. Then the sample covariance of ΔY_t^k is

$$s_{kk} = \frac{1}{T}\sum_{t=1}^{T} \Delta Y_t^{k,2} = \frac{1}{T}\sum_{t=1}^{T}(\alpha^k Z_{1,t-1})^2 + \frac{1}{T}\sum_{t=1}^{T} w_t^{k,2} + \frac{2}{T}\sum_{t=1}^{T} \alpha^k Z_{1,t-1} w_t^k$$

The first to term in (4.2) are positive while the last term in (4.2) converges to zero due to the independence of w_t. Therefore, we conclude that the smaller s_{kk} is, the closer is α^k to zero. To keep only the significant stationary variables in ΔY_t or significant rows in α, we define the pre-screened sets

$$\begin{aligned} \mathbf{B} &= \{k : s_{kk} \geqslant \sigma^2(1+\gamma_T)\} \\ \mathbf{C} &= \{k : s_{kk} < \sigma^2(1+\gamma_T)\} \end{aligned}$$

where $\gamma_T = \gamma\sqrt{\frac{\log m}{T}}$ and γ is the tuning parameter. Besides, denote the cardinality of $\mathbf{B}$ as s_a.

ΔY_t^k, if $k \in \mathbf{C}$, is mainly driven by the white noise term according to the pre-screening rule. Such variables will not provide enough useful information in the canonical correlation analysis. Thus, we only need ΔY_t^k with $k \in \mathbf{B}$ in the RRR step. The next lemma shows that the submatrix $\alpha^{\mathbf{C}}$ of α with row index from $\mathbf{C}$ has insignificant norm.

Lemma 4.3.1. *Under Assumptions 4.2.1 and 4.2.2, for some positive and bounded constants C_0 and C_1, it holds that with probability at least as large as $1 - C_0 m^{-1}$, it holds that*

$$\mathbb{P}\Big(||\alpha^{\mathbf{C}}||_F > C_1 r(\sqrt{\frac{\log m}{T}})^{\frac{d^2}{4-2d}}\Big) \leqslant C_0 m^{-1}$$

Lemma 4.3.1 implies that $\alpha^{\mathbf{B}}$ of α captures almost all the variation due to α.

4.3.2 Pre-screening for cointegrating space

The pre-screening of β as defined in (4.1) relies on the correlation between each ΔY_t^n and Y_{t-1}^k for $k = 1, 2, \cdots, m$ and $n \in \mathbf{B}$, i.e.

$$d_{nk} \quad = \quad \frac{1}{T}\sum_{t=1}^{T}(\Delta Y_{t+1}^n - \Delta Y_t^n)(Y_t^k - Y_{t-1}^k) \tag{4.2}$$

The idea behind the pair-wise correlation defined in (4.2) is that if the k-th row in β has only zero element, the correlation between the increments of Y_t^k and ΔY_{t+1}^n is quite small. To normalize the coefficient of d_{nk} in the last equality of (4.2), we introduce $\tilde{d}_{nk}$ in (4.3).

$$\tilde{d}_{nk} \quad = \quad \frac{d_{nk}}{\frac{1}{T}\sum_{t=1}^{T}\Delta Y_t^{k,2}} \tag{4.3}$$

Such an approach is similar to partial correlation and suffers from endogeneity bias due to the correlation among different ΔY_t^k, i.e., for some k, d_{nk} could be be significantly different from zero even if all the elements in the k-th row of β are all zero. In Theorem 4.3.1 we present the pre-screening criteria for β with statistical properties:

Theorem 4.3.1. *Under Assumptions 4.2.1, 4.2.2 and 4.2.3, define*

$$\mathbf{M}_c = \{1 \leqslant k \leqslant m : \frac{1}{s_a}\sum_{n \in \mathbf{B}\backslash k}\tilde{d}_{nk}^2 > c^2\frac{\log m}{T}\} \tag{4.4}$$

as the subset of relevant Y_{t-1}^k for the reduced rank regression step with $|\mathbf{M}_c| = \hat{K}$. Then it holds that

$$\mathbb{P}\Big(\mathbf{M}_K \subseteq \mathbf{M}_c\Big) \geqslant 1 - O\Big(\frac{rK}{T^{q/2-1}(\log m)^{q/2}} + \frac{rK}{m^c}\Big)$$

where $(\mathbf{E}(|\Delta Y_t^l|^q))^{\frac{1}{q}} < \infty$ for all $l = 1, 2, \ldots, m$.

Remarks: We assume that only K rows in β $(m \times r)$ have non-zero elements. Besides, the marginal effect of each Y_{t-1}^k has r channels on ΔY_t. Both K and r are small compared to m or T. Therefore, the large dimension m would not impose a strict condition on the probability inequality. Moreover, since K and r are small, no very high moment assumption is required at this point.
It should be noted that the pre-selected set $\mathbf{M}_c$ is in general larger than the real $\mathbf{M}_K = \{1 \leqslant k \leqslant m : \beta_k \neq 0\}$. This is due to the fact that by comparing the pair-wise correlations, we can't exclude those variables beyond $\mathbf{M}_K$ but have strong correlations with some variable in $\mathbf{M}_K$. To identify these variables, we could possibly apply iterative penalization approach similar to Ma (2013), which is beyond the scope of this paper.

4.4 Conditional Reduced Rank Regression

In this subsection, we focus only on the ΔY_t^j with $j \in \mathbf{B}$, denoted by ΔY_t^B and Y_{t-1}^j with $j \in \mathbf{M}_c$, denoted by Y_{t-1}^M. The analysis in the pre-screening subsection implies that with large probability, ΔY_t^B and Y_{t-1}^M capture almost all the canonical correlation between ΔY_t and Y_{t-1}. Therefore, in this part, we can focus on the reduced rank regression based on ΔY_t^B and Y_{t-1}^M. From now on, define

$$\begin{aligned} S_{00} &= \frac{1}{T}\sum_{t=1}^{T} \Delta Y_t^B \Delta Y_t^{B\prime} \\ S_{01} &= \frac{1}{T}\sum_{t=1}^{T} \Delta Y_t^B Y_{t-1}^{M\prime} \\ S_{10} &= S_{01}' \\ S_{11} &= \frac{1}{T}\sum_{t=1}^{T} Y_{t-1}^M Y_{t-1}^{M\prime} \end{aligned}$$

By solving the generalized eigenvalue decomposition problem (4.5)

$$S_{10}S_{00}^{-1}S_{01}\hat{h}_i = \hat{\lambda}_i S_{11}\hat{h}_i \tag{4.5}$$

we get the generalized eigenvalues $\hat{\lambda}_1 \geqslant \hat{\lambda}_2 \geqslant \cdots \geqslant \hat{\lambda}_{\hat{K}}$ and the corresponding eigenvectors $\hat{h}_i$, for $i = 1, 2, \dots, \hat{K}$.

The rank can be estimated as

$$\hat{r} = \underset{1 \leqslant j \leqslant [\hat{K}/2]}{\arg\min} \frac{\hat{\lambda}_{j+1}}{\hat{\lambda}_j} \tag{4.6}$$

and the subspace generated by $\{\hat{h}_1, \hat{h}_2, \dots, \hat{h}_{\hat{r}}\}$ is the estimator for the subspace generated by β.

To characterize the accuracy of estimating β with $\hat{h}_j$, we must derive the population counterpart for each term in (4.5) and calculate the distance between the sample covariance matrix and its population counterparts. This requires the transformation from (4.5) to (4.7)

$$Q^M S_{10} S_{00}^{-1} S_{01} Q^{M\prime} D_T^{-1} (Q^{M\prime} D_T^{-1})^{-1} \hat{H} = Q^M S_{11} Q^{M\prime} D_T^{-1} (Q^{M\prime} D_T^{-1})^{-1} \hat{H} \hat{\Lambda} \tag{4.7}$$

or

$$Q^M S_{10} S_{00}^{-1} S_{01} Q^{M\prime} D_T^{-1} \hat{G} = Q^M S_{11} Q^{M\prime} D_T^{-1} \hat{G} \hat{\Lambda} \tag{4.8}$$

where

$$Q^M = \begin{pmatrix} \beta^{M\prime} \\ \alpha_\perp^{M\prime} \end{pmatrix} \qquad D_T = \begin{pmatrix} I_r & 0 \\ 0 & T I_{\hat{K}-r} \end{pmatrix}$$
$$\hat{\Lambda} = diag\{\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_{\hat{K}}\} \qquad \hat{H} = (\hat{h}_1, \hat{h}_2, \dots, \hat{h}_{\hat{K}})$$

with the $\hat{K} \times r$-dimensional matrix $\beta^M(\alpha^M)$ as the submatrix of $\beta(\alpha)$ containing the rows from the set $\mathbf{M}_c$; $\alpha_\perp^M$ is an orthonormal matrix orthogonal to α^M;

Compared to (4.5), (4.7) disentangles the stationary part $Z_{1,t-1} = \beta^{M\prime} \Delta Y_{t-1}^M$ from the unit root process $Z_{2,t-1}^M = \alpha_\perp^{M\prime} Y_{t-1}^M = \sum_{s=0}^{t-1} \alpha_\perp^{M\prime} w_s^M$, which is convenient to derive the population counterparts. For the second, we introduce D_T^{-1} because each element in $\frac{1}{T} \sum_{t=1}^T Z_{2,t-1}^M Z_{2,t-1}^{M\prime}$ would not converge unless divided by one more T. In the final decomposition results, the generalized eigenvalues are not changed from (4.5) to (4.7). Define $\hat{H} = (Q^{M\prime} D_T^{-1})^{-1} \hat{G}$. $\hat{H}_{.1}$ ($\hat{G}_{.1}$) denotes the first r columns in $\hat{H}$ ($\hat{G}$). $\hat{H}_{11}$ and $\hat{H}_{21}$ denotes the first r and the rest $\hat{K} - r$ rows in $\hat{H}_{.1}$ respectively. $\hat{G}_{11}$ and $\hat{G}_{21}$ are defined similarly. Therefore, (4.5) and (4.7) are equivalent. Based on (4.7), we can derive the following results:

Theorem 4.4.1. *Define*

$$\begin{aligned}
\mathcal{L}_0 &= (\beta^{M\prime}\alpha^M)^{-1}(\beta^{M\prime}\frac{\sigma^2}{\sqrt{2}}\mathbf{W}^M\alpha_\perp^{M\prime} + \beta^{M\prime}\Sigma_w^M\alpha_\perp^M) && (4.9)\\
\mathcal{L}_1 &= -\alpha^B\mathcal{L}_0 + \frac{\sigma^2}{\sqrt{2}}\mathbf{W}^{BM}\alpha_\perp^{M\prime} && (4.10)\\
\Sigma_{z1|\Delta Y^B} &= \Sigma_{z1\Delta Y^B}\Sigma_{\Delta Y^B}^{-1}\Sigma_{\Delta Y^B z1} && (4.11)
\end{aligned}$$

where $\mathbf{W}^M = \{\mathbf{W}(i,j) : \quad i,j \in \mathbf{M}_c\}$ *and* $\mathbf{W}^{BM} = \{\mathbf{W}(i,j) : \quad i \in \mathbf{B},\ j \in \mathbf{M}_c\}$ *are subsets of* $\mathbf{W}$*, an* $m \times m$ *dimensional matrix with each element drawn independently and identically from* $N(0,1)$*. Then given Assumptions 4.2.1, 4.2.2 and 4.2.3, it holds that*

$$\begin{aligned}
& ||Q^M S_{10}S_{00}^{-1}S_{01}Q^{M\prime}D_T^{-1} - \begin{pmatrix} \Sigma_{z1|\Delta Y^B} & 0 \\ \mathcal{L}_1'\Sigma_{\Delta Y^B}^{-1}\Sigma_{\Delta Y^B z1} & 0 \end{pmatrix} ||_2 \\
=\ & O_p\Big(\sqrt{\frac{s_a}{T}} + (\log T \vee \sqrt{s_a})\sqrt{\frac{\hat{K} \vee s_a}{T}} + \frac{\sqrt{\hat{K}}}{T}\Big) \\
& ||Q^M S_{11}Q^{M\prime}D_T^{-1} - \begin{pmatrix} \Sigma_{z1} & 0 \\ \mathcal{L}_0' & \alpha_\perp^M \int_0^1 W_s^M W_s^{M\prime} ds \alpha_\perp^{M\prime} \end{pmatrix} ||_2 \\
=\ & O_p\Big(\sqrt{\frac{r}{T}} + \log T\sqrt{\frac{\hat{K} \vee s_a}{T}} + \frac{\sqrt{\hat{K}}}{T}\Big)
\end{aligned}$$

where W_s^M *stands for a* $\hat{K}$*-dimensional standard Brownian motion. By solving the generalized eigenvalue decomposition problem* (4.12) *with* r *non-zero eigenvalues,*

$$\begin{pmatrix} \Sigma_{z1|\Delta Y^B} & 0 \\ \mathcal{L}_1'\Sigma_{\Delta Y^B}^{-1}\Sigma_{\Delta Y^B z1} & 0 \end{pmatrix}\begin{pmatrix} G_{11}^* \\ G_{21}^* \end{pmatrix} = \begin{pmatrix} \Sigma_{z1} & 0 \\ \mathcal{L}_0' & \alpha_\perp^M \int_0^1 W_s^M W_s^{M\prime} ds\alpha_\perp^{M\prime} \end{pmatrix}\begin{pmatrix} G_{11}^* \\ G_{21}^* \end{pmatrix}\Lambda_1^* \tag{4.12}$$

By simple algebra, we can show that

$$G_{21}^* G_{11}^{*-1} = (\alpha_\perp^M \int_0^1 W_s^M W_s^{M\prime} ds\alpha_\perp^{M\prime})^{-1}\Big(\mathcal{L}_1'\Sigma_{\Delta Y^B}^{-1}\Sigma_{\Delta Y^B z1}\Sigma_{z1|\Delta Y^B}^{-1}\Sigma_{z1} - \mathcal{L}_0'\Big)$$

From these results, we can propose the growth rate of $\hat{K}$ and s_a for consistent covariance matrix estimation. The rate of convergence provides an upper bound for the estimation errors, which will play a role in the model selection and estimation results.

From now on we denote $X_{11}, X_{21}, X_{12}, X_{22}$ as the top-left r block, bottom-left $(\hat{K}-r) \times r$ block, top-right $r \times (\hat{K}-r)$ block and bottom-right $(\hat{K}-r)$ block respectively for a matrix $X \in \mathbb{R}^{\hat{K}\times\hat{K}}$.

In order to continue with matrix perturbation theory for this generalized eigenvalue decomposition problem, we want to use Theorem 2.6 in Stewart and Sun (1990). To apply this result we need the following Lemma.

Lemma 4.4.1. *There exist the following decompositions that*

$$\begin{aligned}
\Phi' \underbrace{\begin{pmatrix} \Sigma_{z1|\Delta Y^B} & 0 \\ \mathcal{L}_1' \Sigma_{\Delta Y^B}^{-1} \Sigma_{\Delta Y^B z1} & 0 \end{pmatrix}}_{A} \Upsilon &= \Omega_1 \\
\Phi' \underbrace{\begin{pmatrix} \Sigma_{z1} & 0 \\ \mathcal{L}_0' & \alpha_\perp^M \int_0^1 W_s^M W_s^{M\prime} ds \alpha_\perp^{M\prime} \end{pmatrix}}_{B} \Upsilon &= \Omega_2
\end{aligned}$$

where Ω_1 and Ω_2 are diagonal matrices, with

$$\begin{aligned}
\Phi' &= N_1' \begin{pmatrix} I_r & 0 \\ -\mathcal{L}_1' \Sigma_{\Delta Y^B}^{-1} \Sigma_{\Delta Y^B z1} \Sigma_{z1|\Delta Y^B}^{-1} & I_{\hat{K}-r} \end{pmatrix} \\
\Upsilon &= \begin{pmatrix} I_r & 0 \\ -(\alpha_\perp^M \int_0^1 W_s^M W_s^{M\prime} ds \alpha_\perp^{M\prime})^{-1} (\mathcal{L}_1' \Sigma_{\Delta Y^B}^{-1} \Sigma_{\Delta Y^B z1} \Sigma_{z1|\Delta Y^B}^{-1} + \mathcal{L}_0') & I_{\hat{K}-r} \end{pmatrix} N_1
\end{aligned}$$

for some non-singular matrix N_1. Without loss of generality, we assume that $\Omega_1^2 + \Omega_2^2 = I$, then

$$\begin{aligned}
||N_1||_2 = O_p(1) \qquad & ||N_1^{-1}||_2 = O_p(\sqrt{\hat{K}}) \\
||\Upsilon||_2 = O_p(\sqrt{\hat{K}}) \qquad & ||\Upsilon^{-1}||_2 = O_p(\hat{K})
\end{aligned}$$

Lemma 4.4.1 shows that the regular matrix pair (A, B) is diagonalizable, which is an important prerequisite to study the perturbation of the generalized eigenvalues. With this result, including the L_2 norms of relevant components in the decompositions, we can continue with the consistency of the eigenvalues as shown in Theorem 4.4.2:

Theorem 4.4.2. *Define the diagonal matrix $R = \Omega_1 \Omega_2^{-1}$ and $R_{(j)}$ as its j-th largest element on the diagonal. Under Assumptions 4.2.1, 4.2.2 and 4.2.3, the generalized* *eigenvalues from* (4.5) *satisfy*

$$\max_j |\hat{\lambda}_i - R_{(j)}| = O_p\Big(\hat{K}^{3/2}(\sqrt{\frac{s_a}{T}} + (\log T \vee \sqrt{s_a})\sqrt{\frac{\hat{K} \vee s_a}{T}})\Big)$$

Therefore, with probability larger than $1 - O_p\Big(\hat{K}^{3/2}(\sqrt{\frac{s_a}{T}} + (\log T \vee s_a)\sqrt{\frac{\hat{K} \vee s_a}{T}})\Big)$, *the estimator from* (4.6) *gives the right rank.*

All the generalized eigenvalues from (4.5) are bounded between zero and one and thus would not lead to underestimation problem as in pure factor model such as Lam and Yao (2012). Such a result is supported by simulation results.
In the next step, we can derive the estimation accuracy of the eigenspace. We divide the perturbation analysis for eigenspace into two steps. In the first step (Theorem 4.4.3), we replace the right $(\hat{K} - r)$-columns of $\tilde{A} = Q^M S_{10} S_{00}^{-1} S_{01} Q^{M\prime} D_T^{-1}$ as well as the top-right $r \times (\hat{K} - r)$-block of $\tilde{B} = Q^M S_{11} Q^{M\prime} D_T^{-1}$ by zero to get a block triangular form so that the new matrix pair is also diagonalizable, which is important for the eigenspace perturbation results. In the second step (Theorem 4.4.4), we show that the deleted parts has insignificant effect on the eigenspace estimation compared to the errors in the first step because these parts in $\tilde{A}$ and $\tilde{B}$ converge to zero very fast.

Theorem 4.4.3. *Define* $\tilde{A} = Q^M S_{10} S_{00}^{-1} S_{01} Q^{M\prime} D_T^{-1}$ *and* $\tilde{B} = Q^M S_{11} Q^{M\prime} D_T^{-1}$. *Define*

$$\hat{A} = \begin{pmatrix} \tilde{A}_{11} & 0_{r\times(\hat{K}-r)} \\ \tilde{A}_{21} & 0_{(\hat{K}-r)\times(\hat{K}-r)} \end{pmatrix} \quad \hat{B} = \begin{pmatrix} \tilde{B}_{11} & 0_{r\times(\hat{K}-r)} \\ \tilde{B}_{21} & \tilde{B}_{22} \end{pmatrix}$$

Denote the right r-dimensional right subspace for matrix pair $(\hat{A}, \hat{B})$ *as* $\hat{\Upsilon}_{.1}$*, then under Assumptions 4.2.1, 4.2.2 and 4.2.3, the distance between* $G^*_{.1}$ *and* $\hat{\Upsilon}_{.1}$ *can be described as*

$$||\sin\Theta(G^*_{.1}, \hat{\Upsilon}_{.1})||_2 = O_p\Big(\frac{\hat{K}^{5/2}}{R_{(r)}}(\sqrt{\frac{s_a}{T}} + (\log T \vee \sqrt{s_a})\sqrt{\frac{\hat{K} \vee s_a}{T}})\Big)$$

Remark: Because the subspace can have different matrix representations, the distance between two subspaces must be measured by the ”angle”, which can be referred to Stewart and Sun (1990). Theorem 4.4.3 doesn’t give us the final result about the statistical properties of the right subspace of the matrix pair $(\hat{A}, \hat{B})$ because we omit the part $\tilde{A}_{12}, \tilde{A}_{22}$ and $\tilde{B}_{22}$. Therefore, we need to bound the distance between $\tilde{\Upsilon}_{.1}$, the right eigenspaces of $(\tilde{A}, \tilde{B})$ and $\hat{\Upsilon}_{.1}$ for $(\hat{A}, \hat{B})$. This is summarized in Theorem 4.4.4

Theorem 4.4.4. *Under the same conditions as in Theorem 4.4.3, the distance* *between $\tilde{\Upsilon}_{.1}$ and $\hat{\Upsilon}_{.1}$ caused by the terms $\tilde{A}_{12}, \tilde{A}_{22}$ and $\tilde{B}_{22}$ is characterized by*

$$||\sin\Theta(\tilde{\Upsilon}_{.1}, \hat{\Upsilon}_{.1})||_2 = O_p\Big(\frac{\sqrt{\hat{K}^5}}{TR_{(r)}}\Big)$$

which is much smaller than the distance $||\sin\Theta(G^*_{.1}, \hat{\Upsilon}_{.1})||_2$. Therefore, the relation between the right eigenspace $\hat{H}_{.1}$ in (4.5) *and $\tilde{\Upsilon}_{.1}$ is*

$$\hat{H}_{.1} = \beta^M \tilde{\Upsilon}_{11} + \frac{1}{T}\alpha_{\perp}^M \tilde{\Upsilon}_{21}$$

So far we have derived rank selection consistency in high dimensional VECM and shown the statistical properties of the eigenspace. From the final results in Theorem 4.4.4, we can see that the estimated eigenspace from (4.5) approximates the space generated by β^M plus a disturbances caused by $\alpha_{\perp}^M$. But the perturbation is small after being divided by T, which benefits from the different convergence rate of the non-stationary component Y_{t-1}^M.

4.A Proofs

Proof for Lemma 4.3.1

Proof. For $\gamma_+ > 1$, define

$$\mathbf{B}_- = \{k : \mathbf{E}(\alpha^k Z_{1,t-1})^2 > \sigma^2\gamma_+\gamma\sqrt{\frac{\log m}{T}}\}$$

then with high probability $\mathbf{B}_- \subseteq \mathbf{B}$ because

$$\begin{aligned}
P(\mathbf{B}_- \nsubseteq \mathbf{B}) &= P(\cup_{k\in\mathbf{B}_-}\{s_{kk} < \sigma^2(1+\gamma_T)\}) \\
&\leqslant \sum_{k\in\mathbf{B}_-} P(\{s_{kk}/\sigma_k^2 < \frac{(1+\gamma_T)}{1+\gamma_+\gamma_T}\}) \\
&= \sum_{k\in\mathbf{B}_-} P(\{s_{kk}/\sigma_k^2 - 1 < \frac{(1-\gamma_+)\gamma_T}{1+\gamma_+\gamma_T}\}) \\
&= O_p(me^{-T\frac{(1-\gamma_+)^2\gamma^2\frac{\log m}{T}}{(1+\gamma_+\gamma_T)^2}}) \\
&= O_p(m^{1-(1-\gamma_+)^2\gamma^2})
\end{aligned}$$

While $\mathbf{E}((\alpha^k Z_{1,t-1})^2) = O_p(\lambda_0||\alpha^k||_2^2)$.
Denote λ_j^α and q_j as the j-th eigenvalue and eigenvector of the matrix $\mathbf{E}(Z_{1,t-1}Z'_{1,t-1})$

with $\lambda_1^\alpha \geqslant \cdots \geqslant \lambda_r^\alpha > 0$, then $\mathbf{E}((\alpha^k Z_{1,t-1})^2) = \sum_{j=1}^r \lambda_j^\alpha (\alpha^k q_j)^2 \geqslant \lambda_r^\alpha ||\alpha^k||_2^2$. Define

$$\mathbf{B}_r = \{k : \lambda_r^\alpha ||\alpha^k||_2^2 > \sigma^2\gamma_+\gamma\sqrt{\frac{\log m}{T}}\}$$

and $\mathbf{B}_r \subseteq \mathbf{B}_-$ holds naturally. Thus we have $\mathbf{B}_r \subseteq \mathbf{B}_- \subseteq \mathbf{B}$ and $\mathbf{B}_r^c \supseteq \mathbf{B}^c$.

$$||\alpha^{\mathbf{B}^c}||_F^2 \leqslant ||\alpha^{\mathbf{B}_r^c}||_F^2 \leqslant r\int_V^\infty c^2 v^{-\frac{2}{d}} dv = rc^2\frac{d}{2-d}V^{1-\frac{2}{d}} = rc^2(\frac{1}{c^2\lambda_r^\alpha}\sigma^2\gamma_+\gamma_T)^{\frac{d^2}{4-2d}}$$

where V satisfies $c^2V^{-\frac{2}{d}} = \frac{1}{\lambda_r^\alpha}\sigma^2\gamma_+\gamma_T$. □

Proof for Theorem 4.3.1

Proof. We consider the rank 1 case to simplify the notations. The analysis can be extended to rank r case directly because r is assumed to be finite in this paper. It hods that

$$\begin{aligned} d_{nk} &= \frac{1}{T}\sum_{t=1}^T(\Delta Y_{t+1}^n - \Delta Y_t^n)(Y_t^k - Y_{t-1}^k) \\ &= \alpha^n(\beta'\frac{1}{T}\sum_{t=1}^T \Delta Y_t)\Delta Y_t^k + \frac{1}{T}\sum_{t=1}^T(w_{t+1}^n - w_t^n)\Delta Y_t^k \\ &= \alpha^n(\sum_{l=1}^K \beta_l\frac{1}{T}\sum_{t=1}^T \Delta Y_t^l)\Delta Y_t^k + \frac{1}{T}\sum_{t=1}^T(w_{t+1}^n - w_t^n)\Delta Y_t^k \\ &= \alpha^n\sum_{l=1}^K \beta_l(\frac{1}{T}\sum_{t=1}^T \Delta Y_t^l\Delta Y_t^k) + \frac{1}{T}\sum_{t=1}^T(w_{t+1}^n - w_t^n)\Delta Y_t^k \end{aligned} \tag{4.13}$$

$$\begin{aligned} \tilde{d}_{nk} &= \frac{d_{nk}}{\frac{1}{T}\sum_{t=1}^T \Delta Y_t^{k,2}} \\ &= \alpha^n\sum_{l=1}^K \beta_l(\frac{1}{T}\sum_{t=1}^T \Delta Y_t^l\Delta Y_t^k) + \frac{1}{T}\sum_{t=1}^T(w_{t+1}^n - w_t^n)\Delta Y_t^k \\ &= \alpha^n\left(\beta_k + \sum_{l=1,l\neq k}^K \beta_l\Big(\frac{\frac{1}{T}\sum_{t=1}^T \Delta Y_t^l\Delta Y_t^k}{\frac{1}{T}\sum_{t=1}^T \Delta Y_t^{k,2}}\Big)\right) \\ &+ \frac{\frac{1}{T}\sum_{t=1}^T(w_{t+1}^n - w_t^n)\Delta Y_t^k}{\frac{1}{T}\sum_{t=1}^T \Delta Y_t^{k,2}} \end{aligned} \tag{4.14}$$

According to our assumptions on w_t, we conclude that

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} w_{t+1}^{n}\Delta Y_t^k &\to_p \quad 0 \\ \frac{1}{T}\sum_{t=1}^{T} w_{t}^{n}\Delta Y_t^k &\to_p \quad \begin{cases}\sigma^2 & n=k\\ 0 & n\neq k\end{cases} \end{aligned}$$

Besides,

$$\frac{1}{T}\sum_{t=1}^{T}\Delta Y_t^l \Delta Y_t^k \to_p \mathbf{E}(\Delta Y_t^l \Delta Y_t^k)$$

Therefore, when $k \neq n$, with large probability, we have that

$$\tilde{d}_{nk} \to \alpha_n\Big(\beta_k + \sum_{l=1,l\neq k}^{K}\frac{\mathbf{E}(\Delta Y_t^l \Delta Y_t^k)}{\mathbf{E}(\Delta Y_t^{k,2})}\Big)$$

In the next step, we take the average of $\tilde{d}_{nk}^2$ over $n \in \mathbf{B}\backslash k$, i.e.,

$$\frac{1}{s_a}\sum_{n\in\mathbf{B}\backslash k}\tilde{d}_{nk}^2 \to_p \frac{1}{s_a}\sum_{n\in\mathbf{B}\backslash k}\alpha_n^2\Big(\beta_k + \sum_{l=1,l\neq k}^{K}\frac{\mathbf{E}(\Delta Y_t^l \Delta Y_t^k)}{\mathbf{E}(\Delta Y_t^{k,2})}\Big)^2 \qquad (4.15)$$

For a large number of s_a, the last term in (4.15) converges to zero while the first term is significant for either nonzero β_k or variables which have little correlation with those in the set $\mathbf{B}$.

To derive the tail probability of the distance between $\tilde{d}_{nk}$ and $\beta_k + \sum_{l=1,l\neq k}^{K}\frac{\mathbf{E}(\Delta Y_t^l \Delta Y_t^k)}{\mathbf{E}(\Delta Y_t^{k,2})}$, we need the probability inequality proposed by Liu et al. (2013), i.e.,

$$P\Big(|\epsilon_{jk}| > v\Big) \leqslant \frac{C_1 n}{(nv)^q} + C_2 e^{-C_4 n v^2}$$

Therefore, if we take $v = c\sqrt{\frac{\log m}{T}}$, the probability bound for the tail is

$$\begin{aligned} & rK\Big(\frac{T}{T^q(\log m/T)^{q/2}} + e^{-c\log m}\Big) \\ = & \frac{rK}{T^{q/2-1}(\log m)^{q/2}} + e^{\log r + \log K - c\log m} \end{aligned}$$

□

Proof for Theorem 4.4.1

Proof. Noting that after the Q^M transformation, we have

$$Q^M S_{10} = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^T Z_{1,t-1}\Delta Y_t^{B\prime} \\ \frac{1}{T}\sum_{t=1}^T Z_{2,t-1}^M\Delta Y_t^{B\prime} \end{pmatrix}$$

and thus the reduced rank regression problem is now

$$\begin{pmatrix} \frac{1}{T}Z_{1l}P_B Z_{1l}' & \frac{1}{T^2}Z_{1l}P_B Z_{2l}^{M\prime} \\ \frac{1}{T}Z_{2l}^M P_B Z_{1l}' & \frac{1}{T^2}Z_{2l}^M P_B Z_{2l}^{M\prime} \end{pmatrix}\hat{G} = \begin{pmatrix} \frac{1}{T}Z_{1l}Z_{1l}' & \frac{1}{T^2}Z_{1l}Z_{2l}^{M\prime} \\ \frac{1}{T}Z_{2l}^M Z_{1l}' & \frac{1}{T^2}Z_{2l}^M Z_{2l}^{M\prime} \end{pmatrix}\hat{G}\hat{\Lambda} \tag{4.16}$$

where $Z_{1l} = [Z_{1,0}, Z_{1,1}, \dots, Z_{1,T-1}]$ and $Z_{2l}^M = [Z_{2,0}^M, Z_{2,1}^M, \dots, Z_{2,T-1}^M]$, $P_B = \Delta Y^{B\prime}(\Delta Y^B \Delta Y^{B\prime})^{-1}\Delta Y^B$ with $\Delta Y^B = [\Delta Y_1^B, \dots, \Delta Y_T^B]$.

To derive the population counterparts for each term in the generalized eigenvalue problem in (4.16), we need the large sample performance for each block in the equation. First of all, we can show that

$$\begin{aligned} &||\Sigma_{\Delta Y^B}||_2 < \infty \\ &||\frac{1}{T}\sum_{t=1}^T \Delta Y_t^B \Delta Y_t^{B\prime} - \Sigma_{\Delta Y^B}||_2 = O_p(\sqrt{\frac{s_a}{T}}) \end{aligned}$$

To show the first result, we need to bound the L_2 norm of Σ_{z1} first, i.e.

$$\begin{aligned} \Delta Z_{1,t} &= \beta^{M\prime}\alpha^M Z_{1,t-1} + \beta^{M\prime}w_t^M \\ Z_{1,t} &= (I + \beta^{M\prime}\alpha^M)Z_{1,t-1} + \beta^{M\prime}w_t^M \\ &= \beta^{M\prime}w_t^M + \sum_{j=1}^{\infty}(I + \beta^{M\prime}\alpha^M)^j \beta^{M\prime}w_{t-j}^M \end{aligned}$$

Therefore

$$\Sigma_{z1} = \sigma^2 I_r + \sigma^2 \sum_{j=1}^{\infty}(I + \beta^{M\prime}\alpha^M)^j (I + \beta^{M\prime}\alpha^M)^{j\prime}$$

and

$$\Sigma_{\Delta Y^B} = \alpha^B \Sigma_{z1}\alpha^{B\prime} + \Sigma_w^B$$

Therefore, we can conclude that both Σ_{z1} and $\Sigma_{\Delta Y^B}$ have bounded L_2 norm. Besides, the smallest eigenvalues are both significantly bounded away from zero.

The second result is proved in Theorem 4.B.1.

Moreover, define $S_{\Delta Y^B} = \frac{1}{T}\sum_{t=1}^T \Delta Y_t^B \Delta Y_t^{B\prime}$, then

$$\begin{aligned} & ||S_{\Delta Y^B}^{-1} - \Sigma_{\Delta Y^B}^{-1}||_2 \\ = & ||S_{\Delta Y^B}^{-1}\Big(S_{\Delta Y^B} - \Sigma_{\Delta Y^B}\Big)\Sigma_{\Delta Y^B}^{-1}||_2 \\ \leqslant & ||S_{\Delta Y^B}^{-1}||_2||S_{\Delta Y^B} - \Sigma_{\Delta Y^B}||_2||\Sigma_{\Delta Y^B}^{-1}||_2 \\ = & O_p\Big(||S_{\Delta Y^B} - \Sigma_{\Delta Y^B}||_2||\Sigma_{\Delta Y^B}^{-1}||_2^2\Big) \end{aligned}$$

Similarly, we can show that

$$\begin{aligned} ||\frac{1}{T}\sum_{t=1}^T \Delta Y_t^B Z_{1,t-1}' - \Sigma_{\Delta Y^B z1}||_2 &= O_p(\sqrt{\frac{s_a \vee r}{T}}) \\ ||S_{\Delta Y^B}^{-1}\frac{1}{T}\sum_{t=1}^T \Delta Y_t^B Z_{1,t-1}' - \Sigma_{\Delta Y^B}^{-1}\Sigma_{\Delta Y^B z1}||_2 &= O_p(\sqrt{\frac{s_a}{T}}) \\ ||(\frac{1}{T}\sum_{t=1}^T Z_{1,t-1}\Delta Y_t^{B\prime})S_{\Delta Y^B}^{-1}(\frac{1}{T}\sum_{t=1}^T \Delta Y_t^B Z_{1,t-1}') - \Sigma_{z1|\Delta Y^B}||_2 &= O_p(\sqrt{\frac{s_a}{T}}) \end{aligned}$$

where $\Sigma_{z1|\Delta Y^B} = \Sigma_{z1\Delta Y^B}\Sigma_{\Delta Y^B}^{-1}\Sigma_{\Delta Y^B z1}$.

In the next we can discuss each block as follows:
Left: Top-left

$$||\frac{1}{T}Z_{1l}P_B Z_{1l}' - \Sigma_{z1\Delta Y^B}\Sigma_{\Delta Y^B}^{-1}\Sigma_{\Delta Y^B z1}||_2 = O_p(\sqrt{\frac{s_a}{T}})$$

is a direct result from the argument before since all the terms are stationary.
Left: Bottom-left Different from Top-Left block, the bottom-left block contains a non-stationary part in

$$\begin{aligned} & \frac{1}{T}\sum_{t=1}^T \Delta Y_t^B Z_{2,t-1}^{M\prime} \\ = & \alpha^B \frac{1}{T}\sum_{t=1}^T Z_{1,t-1}Z_{2,t-1}^{M\prime} + \frac{1}{T}\sum_{t=1}^T w_t^B Z_{2,t-1}^{M\prime} \end{aligned}$$

To study the term $\frac{1}{T}\sum_{t=1}^T Z_{1,t-1}Z_{2,t-1}^{M\prime}$ we rely on two expressions of $\sum_{t=1}^T \Delta Z_{1,t}Z_{2,t-1}^{M\prime}$, i.e., on one hand

$$
\begin{array}{rcl}
\sum_{t=1}^{T} \Delta Z_{1,t} Z_{2,t-1}^{M\prime} & = & \beta^{M\prime} \sum_{t=1}^{T} \Delta Y_t^M Z_{2,t-1}^{M\prime} \\
& = & -\sum_{t=1}^{T} Z_{1,t} \Delta Z_{2,t}^{M\prime} + Z_{1,T} Z_{2,T-1}^{M\prime} - Z_{1,0} Z_{2,0}^{M\prime} \\
& = & -\sum_{t=1}^{T} Z_{1,t} w_t^{M\prime} \alpha_{\perp}^{M\prime} + Z_{1,T} Z_{2,T-1}^{M\prime} - Z_{1,0} Z_{2,0}^{M\prime} \\
& = & -\sum_{t=1}^{T} (\beta^{M\prime} \alpha^M + I_r) Z_{1,t-1} w_t^{M\prime} \alpha_{\perp}^M - \sum_{t=1}^{T} \beta^{M\prime} w_t^M w_t^{M\prime} \alpha_{\perp}^M \\
& + & Z_{1,T} Z_{2,T-1}^{M\prime} - Z_{1,0} Z_{2,0}^{M\prime}
\end{array}
$$

On the other hand,

$$
\sum_{t=1}^{T} \Delta Z_{1,t} Z_{2,t-1}^{M\prime} = \sum_{t=1}^{T} \beta^{M\prime} \alpha^M Z_{1,t-1} Z_{2,t-1}^{M\prime} + \sum_{t=1}^{T} \beta^{M\prime} w_t^M Z_{2,t-1}^{M\prime}
$$

Therefore, we get

$$
||\beta^{M\prime} \alpha^M \frac{1}{T} \sum_{t=1}^{T} Z_{1,t-1} Z_{2,t-1}^{M\prime} + \frac{1}{T} \sum_{t=1}^{T} \beta^{M\prime} w_t^M Z_{2,t-1}^{M\prime} + \beta^{M\prime} \Sigma_w^M \alpha_{\perp}^M ||_F = O_p(\sqrt{\frac{\hat{K}}{T}}) \tag{4.17}
$$

Another partial sum term to be approximated in (4.17) is $\frac{1}{T} \sum_{t=1}^{T} \beta^{M\prime} w_t^M Z_{2,t-1}^{M\prime}$ which is generated by $\frac{1}{T} \sum_{t=1}^{T} w_t W_{t-1}^{\prime}$ with $W_{t-1} = \sum_{s=0}^{t-1} w_s$ and $w_t \sim_{i.i.d} \mathbf{N}(0, \sigma^2 I_m)$. According to Theorem 4.B.2, we conclude that

$$
||\frac{1}{T} \sum_{t=1}^{T} Z_{1,t-1} Z_{2,t-1}^{M\prime} + (\beta^{M\prime} \alpha^M)^{-1} \Big(\beta^{M\prime} \frac{\sigma^2}{\sqrt{2}} \mathbf{W}^M \alpha_{\perp}^{M\prime} + \beta^{M\prime} \Sigma_w^M \alpha_{\perp}^M \Big) ||_2 = O_p(\log T \sqrt{\frac{\hat{K}}{T}}) \tag{4.18}
$$

Therefore

$$\begin{aligned}
& \frac{1}{T}\sum_{t=1}^{T}\Delta Y_t^B Z_{2,t-1}^{M\prime} \\
= \quad & \alpha^B \frac{1}{T}\sum_{t=1}^{T} Z_{1,t-1}Z_{2,t-1}^{M\prime} + \frac{1}{T}\sum_{t=1}^{T} w_t^B Z_{2,t-1}^{M\prime} \\
\rightarrow \quad & -\alpha^B\Big((\beta^{M\prime}\alpha^M)^{-1}(\beta^{M\prime}\frac{\sigma^2}{\sqrt{2}}\mathbf{W}^M\alpha_\perp^{M\prime} + \beta^{M\prime}\Sigma_w^M\alpha_\perp^M)\Big) + \frac{\sigma^2}{\sqrt{2}}\mathbf{W}^{BM}\alpha_\perp^{M\prime}
\end{aligned}$$

It should be noted that the L_2 norm of $\mathbf{W}$ depends on its dimension, i.e.,

$$\begin{aligned}
||\mathbf{W}^M||_2 &= O_p(\sqrt{\hat{K}}) \\
||\mathbf{W}^{BM}||_2 &= O_p(\sqrt{\hat{K} \vee s_a})
\end{aligned}$$

according to Theorem 2.13 of Chapter 8 in Johnson and Lindenstrauss (2001). Thus the $\mathcal{L}_0$ and $\mathcal{L}_1$ defined in (4.9) and (4.10) satisfy

$$||\mathcal{L}_0||_2 = O_p(\sqrt{\hat{K}}) \qquad ||\mathcal{L}_1||_2 = O_p(\sqrt{\hat{K} \vee s_a})$$

Therefore,

$$\begin{aligned}
& ||(\frac{1}{T}\sum_{t=1}^{T} Z_{2,t-1}^M \Delta Y_t^{B\prime})S_{\Delta Y^B}^{-1}(\frac{1}{T}\sum_{t=1}^{T}\Delta Y_t^B Z_{1,t-1}^\prime) - \mathcal{L}_1^\prime \Sigma_{\Delta Y^B}^{-1}\Sigma_{\Delta Y^B z1}||_2 \\
\leqslant \quad & ||\frac{1}{T}\sum_{t=1}^{T} Z_{2,t-1}^M \Delta Y_t^{B\prime} - \mathcal{L}_1^\prime||_2 ||S_{\Delta Y^B}^{-1}(\frac{1}{T}\sum_{t=1}^{T}\Delta Y_t^B Z_{1,t-1}^\prime)||_2 \\
+ \quad & ||\mathcal{L}_1||_2 ||S_{\Delta Y^B}^{-1}(\frac{1}{T}\sum_{t=1}^{T}\Delta Y_t^B Z_{1,t-1}^\prime) - \Sigma_{\Delta Y^B}^{-1}\Sigma_{\Delta Y^B z1}||_2 \\
= \quad & O_p\Big(\log T\sqrt{\frac{\hat{K}}{T}} + \sqrt{\frac{(\hat{K} \vee s_a)s_a}{T}}\Big) = O_p\Big((\log T \vee \sqrt{s_a})\sqrt{\frac{\hat{K} \vee s_a}{T}}\Big)
\end{aligned}$$

Left- Right Panel According to previous analysis, we can easily show that

$$\begin{aligned}
||\frac{1}{T^2}Z_{1l}P_B Z_{2l}^{M\prime}||_2 &= O_p(\frac{\sqrt{\hat{K}}}{T}) \\
||\frac{1}{T^2}Z_{2l}^M P_B Z_{2l}^{M\prime}||_2 &= O_p(\frac{\hat{K}}{T})
\end{aligned}$$

Right

$$\begin{aligned}
||\frac{1}{T}\sum_{t=1}^{T} Z_{1,t-1}Z_{1,t-1}' - \Sigma_{z1}||_2 &= O_p(\sqrt{\frac{r}{T}}) \\
||\frac{1}{T}\sum_{t=1}^{T} Z_{1,t-1}Z_{2,t-1}^{M\prime} - \mathcal{L}_0||_2 &= O_p(\log T\sqrt{\frac{\hat{K}}{T}}) \\
||\frac{1}{T^2}\sum_{t=1}^{T} Z_{2,t-1}^{M}Z_{1,t-1}'||_2 &= O_p(\frac{\sqrt{\hat{K}}}{T}) \\
||\frac{1}{T^2}\sum_{t=1}^{T} Z_{2,t-1}^{M}Z_{2,t-1}^{M\prime} - \alpha_{\perp}^{M}\int_0^1 W_s^M W_s^{M\prime} ds \alpha_{\perp}^{M\prime}||_2 &= O_p(\log T\sqrt{\frac{\hat{K}}{T}})
\end{aligned}$$

□

Proof for Lemma 4.4.1

Proof. A and B are both block lower-diagonal matrices with the blocks on the diagonal symmetric. Besides, the bottom-right block of A is zero but that of B is positive-definite. We conclude that there exist matrices $P_1 = A_{21}A_{11}^{-1}$ and $P_2 = B_{22}^{-1}(P_1 B_{11} + B_{21})$ of dimension $(\hat{K}-r)\times r$ thus that

$$\begin{aligned}
\begin{pmatrix} I_r & 0 \\ P_1 & I_{\hat{K}-r} \end{pmatrix}\begin{pmatrix} A_{11} & 0 \\ A_{21} & 0 \end{pmatrix}\begin{pmatrix} I_r & 0 \\ P_2 & I_{\hat{K}-r} \end{pmatrix} &= \begin{pmatrix} A_{11} & 0 \\ 0 & 0 \end{pmatrix} = A_d \\
\begin{pmatrix} I_r & 0 \\ P_1 & I_{\hat{K}-r} \end{pmatrix}\begin{pmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{pmatrix}\begin{pmatrix} I_r & 0 \\ P_2 & I_{\hat{K}-r} \end{pmatrix} &= \begin{pmatrix} B_{11} & 0 \\ 0 & B_{22} \end{pmatrix} = B_d
\end{aligned}$$

and $||P_1||_2 = O(\sqrt{\hat{K}})$, $||P_2||_2 = O(\sqrt{\hat{K}})$.

Moreover, due to the fact that $A_{11},\ B_{11},\ B_{22}$ are symmetric, there exist invertible matrices N_1 and diagonal matrix Λ thus that

$$\begin{aligned}
N_1'\begin{pmatrix} A_{11} & 0 \\ 0 & 0 \end{pmatrix}N_1 &= \Omega_1 \\
N_1'\begin{pmatrix} B_{11} & 0 \\ 0 & B_{22} \end{pmatrix}N_1 &= \Omega_2
\end{aligned}$$

with $||N||_2 = O_p\Big(1\Big), ||N^{-1}||_2 = O_p\Big(\sqrt{\hat{K}}\Big)$. Note that the Crawford number for the matrix pair (A_d, B_d) is significantly different from zero. Therefore, we can omit this term from the analysis. Let $\Upsilon = N_1 \begin{pmatrix} I_r & 0 \\ P_2 & I_{\hat{K}-r} \end{pmatrix}$, thus

$$||\Upsilon||_2 = O_p\Big(\sqrt{\hat{K}}\Big) \qquad ||\Upsilon^{-1}||_2 = O_p\Big(\hat{K}\Big)$$

□

Proof for Theorem **4.4.2**

Proof. The perturbation result for the eigenvalues are direct result from Theorem 2.6 of Stewart and Sun (1990) based on Lemma 4.4.1. Only the first largest $R_{(j)}$ are nonzero and all the others are zero. Therefore, $\hat{\lambda}_{r+1}$ is close to zero and thus $\frac{\hat{\lambda}_{r+1}}{\hat{\lambda}_r}$ gives the first ratio jump in this test. □

Proof for Theorem **4.4.3**

Proof. We can transform matrix pair $(\hat{A}, \hat{B})$ in a similar matrix as Lemma 4.4.1 and get the same upper bounds for the L_2 norm of relevant components. Then Theorem 5.3 of Li (1994) can be applied to get the result. □

Proof for Theorem **4.4.4**

Proof. The effect of small perturbation from the matrix pair on the right eigenspace is derived in Theorem 2.14 of Stewart and Sun (1990). The result can be derived by noting that $||\tilde{A}_{12}||_2 = O(\frac{\sqrt{\hat{K}}}{T})$, $||\tilde{A}_{22}||_2 = O_p(\frac{\hat{K}}{T})$ and $||B_{12}||_2 = O_p(\frac{\sqrt{\hat{K}}}{T})$. Simple linear algebra will help derive the results. □

4.B Auxiliary Technical Results

Theorem 4.B.1. *Under Assumption 4.2.1 and w_t follows i.i.d. $N(0, \sigma^2 I_m)$ process, it holds that*

$$||\frac{1}{T}\sum_{t=1}^{T} Z_{1,t-1}Z'_{1,t-1} - \Sigma_{z1}||_2 = O_p(\sqrt{\frac{r}{T}})$$
$$||\frac{1}{T}\sum_{t=1}^{T} \Delta Y_t^B \Delta Y_t^{B\prime} - \Sigma_{\Delta Y^B}||_2 = O_p(\sqrt{\frac{s_a}{T}})$$

Proof. Define $v_{1,t} = w_t$, then $v_{1,t}$ follows i.i.d. $N(0, I_r)$ and

$$||\frac{1}{T}\sum_{t=1}^{T} v_{1,t}v_{1,t}' - I_r||_2 = O(\sqrt{\frac{r}{T}})$$

where the probability bound is derived in Proposition D.1 in the Supplement to Ma (2013). Besides, by expressing $Z_{1,t}$ as MA representation

$$Z_{1,t} = \sum_{j=0}^{\infty}(\beta'\alpha + I_r)^j v_{1,t-j}$$

we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} Z_{1,t}Z_{1,t}' &= \sum_{j=0,k=0}^{\infty} (\beta'\alpha + I_r)^j \Big(\frac{1}{T}\sum_{t=1}^{T} v_{1,t-j}v_{1,t-k}'\Big)(\beta'\alpha + I_r)^{k\prime} \\
&= \sum_{j=0}^{\infty}(\beta'\alpha + I_r)^j \Big(\frac{1}{T}\sum_{t=1}^{T} v_{1,t-j}v_{1,t-j}'\Big)(\beta'\alpha + I_r)^{j\prime} \\
&+ \sum_{j=0}^{\infty}\sum_{h=-\infty,h\neq 0}^{\infty} (\beta'\alpha + I_r)^j \Big(\frac{1}{T}\sum_{t=1}^{T} v_{1,t-j}v_{1,t-j-h}'\Big)(\beta'\alpha + I_r)^{j+h\prime}
\end{aligned}$$

The first term in the last equality converges to $\sum_{j=0}^{\infty}(\beta'\alpha + I_r)^j(\beta'\alpha + I_r)^{j\prime}$ and the second term goes to zero. According to standard results from matrix perturbation theory,

$$||\frac{1}{T}\sum_{t=1}^{T} v_{1,t-j}v_{1,t-k}'||_2 \leqslant \frac{1}{T}\sum_{t=1}^{T} ||v_{1,t-j}v_{1,t-k}'||_2 = O_p(\sqrt{\frac{r}{T}}) \tag{4.19}$$

Therefore, we have the first result in the Theorem. The second result can be proved in the similar manner. □

Theorem 4.B.2. *By strong invariance principle, each component* $\frac{1}{T}\sum_{t=1}^{T} w_t^i W_{t-1}^j$ *can be strongly approximated by a Brownian motion. Besides, the elements in* $w_t W_{t-1}'$ *are independent of one another.*

$$|\frac{1}{T}\sum_{t=1}^{T} w_t^i W_{t-1}^j - \frac{\sigma^2}{\sqrt{2}}\mathbf{W}_{ij}(1)| = O_{a.s.}(\frac{\log T}{\sqrt{T}})$$

Proof.

$$\begin{aligned}\mathbf{E}\Big((\frac{1}{T}\sum_{t=1}^{T} w_t^i W_{t-1}^j)^2\Big) &= \frac{1}{T^2}\sum_{t=1}^{T}\mathbf{E}(w_t^{i2} W_{t-1}^{j2}) \\ &= \frac{1}{T^2}\sigma^2\sum_{t=1}^{T}\mathbf{E}(W_{t-1}^{j2}) \to \frac{1}{2}\sigma^4\end{aligned}$$

The cross-terms disappear because of the independence across time, i.e. for any $h > 0$, it holds that

$$\mathbf{E}(w_t^i W_{t-1}^j w_{t-h}^i W_{t-1-h}^j) = \mathbf{E}(W_{t-1}^j w_{t-h}^i W_{t-1-h}^j \mathbf{E}(w_t^i|\mathcal{F}_{t-1})) = 0$$

The independence for different i, j can be derived in a similar manner

$$\begin{aligned}\mathbf{E}(w_t^i W_{t-1}^j w_t^h W_{t-1}^k) &= \mathbf{E}\Big(W_{t-1}^j W_{t-1}^k E(w_t^i w_t^h|\mathcal{F}_{t-1})\Big) \\ &= \begin{cases}0 & i \neq h \\ \sigma^2 E(W_{t-1}^j W_{t-1}^k) & i = h\end{cases} \\ E(W_{t-1}^j W_{t-1}^k) &= 0 \quad \text{if } j \neq k\end{aligned}$$

This implies that $\frac{1}{T}\sum_{t=1}^{T} w_t^i W_{t-1}^j$ and $\frac{1}{T}\sum_{t=1}^{T} w_t^h W_{t-1}^k$ are independent of each other as long as $i \neq h$ or $h \neq k$. □

5 Time-varying Limit Order Book Networks

5.1 Introduction

Advancements in trading technologies allow an extremely rapid placement of buy and sell orders. These rapid-fire trading algorithms can make decisions in milliseconds. The dynamic changes of the high frequency (HF) limit order book (LOB) gives us vital insights into the market behavior. In an LOB shown in Figure 5.1, the order book contains a quantity of limit orders and the corresponding price at which you would issue a "buy" or "sell" limit order. When an investor places an order to purchase or sell a stock, there are two fundamental execution options: place the order "at market" or "at limit." The market ones are orders of purchase or sale at the best available quote. On the other hand the limit orders are not immediately executed since they are placed at a quote which is less favorable than the best quote, e.g. the second level bid/ask order. The schematic representation of an LOB reflects the local decisions and interactions between thousands of investors, and thus generates a high dimensional dynamic and complex system. Insights into this highly dynamic LOB is therefore vital for pricing of assets, but requires skillful dimension reduction techniques in combination with generalized impulse response analysis.

The limit order book has been analyzed in a variety of ways, theoretical analysis of limit orders include Parlour and Seppi (2003), Foucault et al. (2005), Roşu (2009) etc. Empirical examples are Handa et al. (2003) and Bloomfield et al. (2005). These pieces of literature provide useful characterizations of limit orders, and discuss in detail the evolution of liquidity in an LOB market. Kavajecz and Odders-White (2004) suggests that limit orders may, in part, be informative about pockets of future liquidity. However empirical evidence on the actual market impact of limit order placements across stocks is virtually not existent, many questions of interest to regulators and traders are unsolved: i) How does the order flows interact with price dynamics, and further affect the market behavior? ii) Are the impacts on price responding to incoming ask and bid market/limit orders widely symmetric? iii) If not symmetric, how does the heterogeneous market impact caused by bid and ask order for various stocks affect the whole market? iv) How to measure the impact of market/limit order quantitatively? To address the arising questions, in this paper

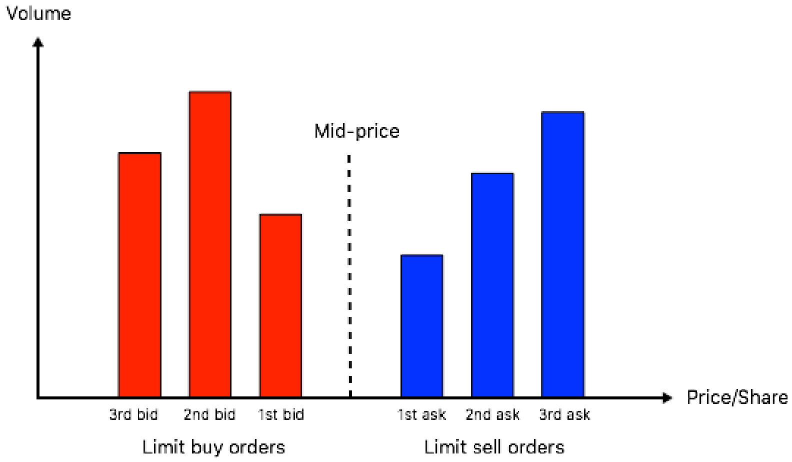


Figure 5.1. A simplified example of the three level LOB, with market order and first two levels of limit orders.

we conduct a comprehensive study on the interaction among price, bid and ask order sizes. LOB provides a more complicated scenario that inspires us to construct a high-dimensional object using both price and several levels of depth of order sizes with historical order flow. Of particular interest is vast directed network analysis based on the constructed high-dimensional object. The underlying assumption is that there exists a sparse representation of the data. This may help us to understand how information is impounded into price. For instance, the orders posted on the selected order levels that induce significant price impact would be treated as price drivers. In this way, investors' decision-making can be addressed by making trading price driven by order flows. The motivation to construct a network of LOB stems from the lack of both theoretical setup and empirical support.

To do so the vector autoregressive (VAR) model is without doubt one of the most useful tools that allows us to capture in a simple fashion their dynamic evolution. However it imposes challenges of high dimensionality when we incorporating a variety of time series, particularly where the vector observed at each time is high dimensional relative to the time period. Researchers have developed various penalized estimators to filter out less relevant variables, key papers are on the Lasso Tibshirani (1996), SCAD Fan and Li (2001), adaptive Lasso Zou (2006), elastic net Zou and Hastie (2005), Dantzig selector Candes and Tao (2007). This paper is different from this structure of thoughts since it focuses on network connectivity, which is derived from generalized impulse response function. There has been a large literature discussing sparse VAR estimation through different penalty terms. For instance, Negahban and Wainwright (2011) imposed sparse dependence assumption on the transition matrix of VAR model and studied the theoretical properties. Kock and Callot (2015) discussed theoretical properties of Lasso and adaptive Lasso in

VAR model that may reveal the correct sparsity pattern asymptotically. Basu and Michailidis (2015) investigated theoretical properties of Lasso-type estimators for high-dimensional Gaussian processes. Wu and Wu (2016) studied the systematic theory for high-dimensional linear models with correlated errors. The Lasso-type estimators penalize the regression coefficients with the model size via a shrinkage procedure. Belloni et al. (2012) and Belloni and Chernozhukov (2013) studied the post-model selection estimator that apply OLS to the first-step penalized estimators to alleviate shrinkage bias.

Diebold and Yılmaz (2014) proposed connectedness measures built from generalized forecast error variance decomposition (GFEVD) based on VAR systems, where the GFEVD is developed by Pesaran and Shin (1998) and Koop et al. (1996) with an intrinsic appeal to order-invariance. However the contributions of shares of forecast error variation in various locations do not add to unity, and it is restricted to Gaussian innovations. To solve this, we use the LN-GFEVD that has been recently proposed by Lanne and Nyberg (2016). The LN-GFEVD is economically interpretable, and can be implemented to both linear (Gaussian and non-Gaussian) or nonlinear models. To keep the sparsity structure of high-dimensional VAR estimator, we apply a bootstrap-based method rather than a moving-average (MA) transformation which is often done in fixed dimensional cases. In summary a new connectedness table is obtained, where the directed connectedness "from" and "to" are associated with the new forecast error variation for specific order book across various stocks when the arising shocks transmit from one stock to the others. This paper contributes to network construction through high-dimensional VAR estimation, the resulting connectedness table facilitates convenient interpretation. At the same time, a parsimonious algorithm without MA transformation can help to improve the accuracy of final connectedness estimator.

We progress by focusing on the dynamics of LOB networks and their evolution. First, we find that the network involving the trading volumes is a better measure of the stock connectedness with the finance sector dominating the market in the sense of having a stronger influence on the others. Second, financial stocks are size-dominated, their price patterns are highly related to the market trading activity. The impact caused by ask and bid orders are statistically significant, substantial in size and significantly asymmetric. In particular, the NASDAQ market is more sensitive to the market sell pressure. Third, we investigate the LOB trading activity and find significant own-price and cross-price market impact. Moreover, we are able to identify the significant market impact caused by the arrival of a large market/limit order, and several robust risk transmission channels. Overall, our findings on the time-varying LOB networks yield a better understanding of market behavior.

The rest of the paper is organized as follows. Section 5.2 introduces NASDAQ LOB market and the non-synchronous LOB data, we then elaborate the data preparation based on volume-synchronization algorithm. In Section 5.3 we present the theoretical framework for high-dimensional VAR estimation, and construct the connectedness estimator based on our setting. The empirical results of time-varying network are illustrated in Section 5.4. Section 5.5 measures price dynamics under uncertainty shocks. Section 5.6 concludes, while more technical details are relegated to the Appendix.

5.2 Description of the Market and Data Preparation

5.2.1 NASDAQ Limit Order Book Market

Our sample consists of intraday trading data for selected NASDAQ stocks for the sample period spanning 1st, June 2016 to 30th, July of 2016. These data come from the LOBSTER academic data, which is powered by NASDAQ's historical TotalView using very detailed event information.

The basic structure of LOB is shown in Figure 5.2. The sample file has one time-stamped record for every order entered for each stock throughout the trading day. Trades are time stamped up to the nanosecond and signed to indicate whether they were initiated by a buyer or seller by the "Direction" ticker, i.e. sell trade direction are set to '-1' and buy trade direction are set to '1'. The ticker of "Event Type" indicates the trading type, for example, 1: Submission of a new order, 2: Cancellation (partial deletion of a order order), 3: Deletion (total deletion of a market/limit order), 4: Execution of a visible limit order, 5: Execution of a hidden limit order etc. Another important feature of this dataset is that each quote has been associated with trading information and limit order book. To be more specific, the k-th row in the "message" file (upper panel of Figure 5.2) describes the limit order event causing the change in the limit order book from line $k-1$ to line k in the "orderbook" file (lower panel).

The sample is stratified by market capitalization and industry sector. The industry breakdown of NASDAQ market is technology of 45.38%, Health Care of 11.43% and Financials of 8.42% (as of 23.02.2018). We consider a sample portfolio with 9 assets listed in Table 5.1, together with their market order and first two levels of limit orders, which attract the majority of trading activity, therefore becoming our research interest.

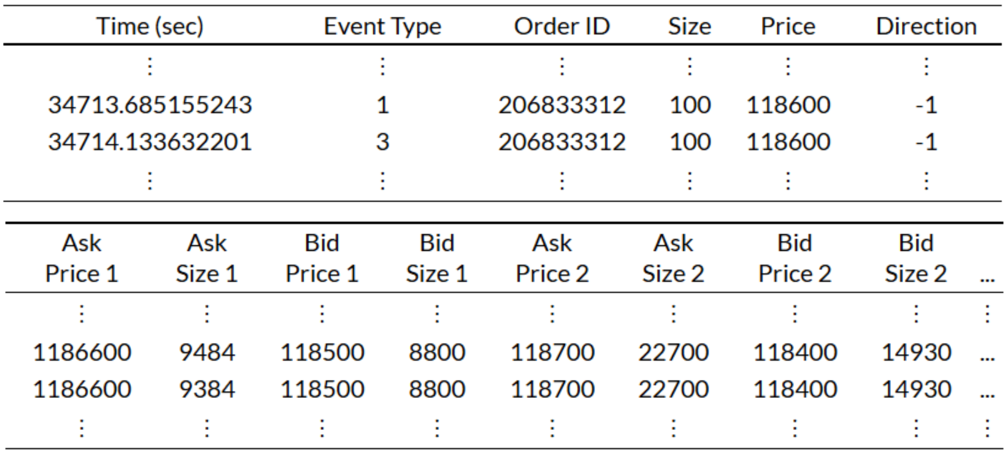
We present the summary statistics of sample dataset in Table 5.2. The data is collected for the normal trading day involving both visible and hidden orders, which

Industry	Stock	Company	MktCap (billion $)
Technology	IBM	International Business Machines Corp.	171.72
	MSFT	Microsoft Corporation	499.35
	T	AT&T Inc.	257.53
Healthcare	JNJ	Johnson & Johnson	328.91
	PFE	Pfizer Inc.	206.69
	MRK	Merck & Co. Inc.	181.56
Finance	JPM	JP Morgan Chase & Co.	326.04
	WFC	Wells Fargo & Company	293.39
	C	Citigroup Inc.	168.06

Table 5.1. Sample data. MktCap is the market capitalization by Feb 25th, 2017.

	NumObs	AvgTrd	AvgAP1	AvgBP1	AvgAS1
	$(*10^3)$	$(*10^3)$	(in $)	(in $)	(100 shrs)
IBM	118.25	5.82	153.07	153.04	1.92
MSFT	584.55	25.91	52.28	52.26	24.19
T	223.45	6.67	38.75	38.74	36.36
JNJ	172.77	8.17	113.99	113.98	4.11
PFE	427.51	12.49	34.83	34.82	41.96
MRK	188.84	5.82	56.70	56.68	7.43
JPM	414.35	11.49	65.48	65.46	9.47
WFC	275.29	10.91	50.90	50.89	18.02
C	472.90	12.19	46.82	46.81	14.19
	AvgBS1	AvgAS2	AvgBS2	AvgAS3	AvgBS3
	(100 shrs)	(100 shrs)	(100 shrs)	(100 shrs)	(100 shrs)
IBM	2.17	1.95	2.26	2.09	2.26
MSFT	24.53	28.12	31.06	33.90	35.37
T	33.76	43.63	41.96	55.53	63.67
JNJ	3.62	5.86	4.44	7.74	4.90
PFE	42.29	48.07	48.09	50.94	55.68
MRK	7.36	14.34	11.30	24.20	13.87
JPM	9.45	13.10	11.82	17.41	15.09
WFC	17.01	20.68	17.72	23.58	19.05
C	12.97	18.58	16.48	22.23	19.60

Table 5.2. Summary statistics of selected stocks. $NumObs$ denotes the average number of observation. $AvgTrd$ is the average number of execution trades of a market/limit order. $AvgAP1$ gives the average ask price for the first order, and $AvgAS1$ represents the corresponding ask size.



Time (sec)	Event Type	Order ID	Size	Price	Direction
⋮	⋮	⋮	⋮	⋮	⋮
34713.685155243	1	206833312	100	118600	-1
34714.133632201	3	206833312	100	118600	-1
⋮	⋮	⋮	⋮	⋮	⋮

Ask Price 1	Ask Size 1	Bid Price 1	Bid Size 1	Ask Price 2	Ask Size 2	Bid Price 2	Bid Size 2	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1186600	9484	118500	8800	118700	22700	118400	14930	...
1186600	9384	118500	8800	118700	22700	118400	14930	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 5.2. Structure of LOBSTER data

run from 9:30 a.m to 4 p.m ET. To avoid erratic effects during the market opening and closure, our sample period covers only the continuous trading periods between 9:45 and 16:00.

The main challenge in dealing with HFT data is the presence of microstructure noise arising from market frictions, where the noise-induced bias at very high sampling frequencies contaminates the observed price. Whereas infrequent sampling frequency leads to imprecise estimates, optimal sampling frequency is needed to acquire bias-variance tradeoff, see Bandi and Russell (2006), Aït-Sahalia et al. (2005), Bandi and Russell (2008). Here we implement the pre-averaging approach to exclude the impact of microstructure noise, technical details can be found in Appendix 5.A.

5.2.2 Volume synchronization Algorithm

This section is devoted to the data preparation procedure by involving the order flows. We propose an algorithm that achieves volume synchronization for high-dimensional statistical setting.

As we know, the market order gets transacted at whatever price in that market, while the limit order specify the price at which to execute the order. For larger orders placed in the market, it takes longer to fill and can actually move the market on their own. In contrast to a moderate time interval for price to reduce the microstructure noise, the time interval for trading volumes should be small enough to capture the large orders submitted by the market trader. Considering the facts, we propose a trading volume measure, size intensity, $\tilde{S}_{t_j}$ that captures the trading crowd that provides a substantial amount of liquidity at the quotes,

$$\tilde{S}_{t_j} = S_{t_j}(t_{j+1} - t_j) \tag{5.1}$$

where t_j denotes the time stamp of jth LOB, S_{t_j} is the corresponding tick size at t_j. By size intensity can be summed up over a given moderate time interval and therefore, matched with returns.

In the following we shall illustrate how to explicitly prepare the raw HF data. For ease of illustration, the volume synchronization algorithm can be divided into four steps,

Step 1: Set equally-spaced k time intervals starting at time T_0

$$T_0 + k\Delta T, \quad k = 0, 1, 2, \dots, K$$

Step 2: Define the price and size at time $T_0 + k\Delta T$ as

$$\begin{aligned} \tilde{P}_{T_0+k\Delta T} &= P_{t_m},\ t_m = \max\{t_j;\ t_j \leqslant T_0 + k\Delta T\} \\ \tilde{S}_{T_0+k\Delta T} &= \sum_{T_0+(k-1)\Delta T \leqslant t_j \leqslant T_0+k\Delta T} S_{t_j}(t_{j+1} - t_j) \end{aligned}$$

the size variables denoted as $\tilde{S}_{T_0+k\Delta T}$ are the size intensity measure (5.1).

Step 3: Compute the changes of the log values,

$$\begin{aligned} \Delta p_{T_0+k\Delta T} &= \log \tilde{P}_{T_0+k\Delta T} - \log \tilde{P}_{T_0+(k-1)\Delta T} \\ \Delta s_{T_0+k\Delta T} &= \log \tilde{S}_{T_0+k\Delta T} - \log \tilde{S}_{T_0+(k-1)\Delta T} \end{aligned}$$

Step 4: Pre-averaging both $\Delta p_{T_0+k\Delta T}$ and $\Delta s_{T_0+k\Delta T}$ to remove microstructure noise,

$$\begin{aligned} \Delta \tilde{p}_{T_0+k\Delta T} &= \sum_{j=0}^{J} g_j \Delta p_{T_0+j\Delta T} \\ \Delta \tilde{s}_{T_0+k\Delta T} &= \sum_{j=0}^{J} g_j \Delta s_{T_0+j\Delta T} \end{aligned}$$

where $g_j \geqslant 0$ and $\sum_{j=0}^{J} g_j = 1$, the details are in Appendix 5.A.

Preparing data in this way alleviates microstructure noise, matches the price to the size in a moderate interval and solves the problem of non-synchronicity.

For each stock, we take the mid price on the first level and the corresponding bid and ask sizes on the first three levels, i.e. market order, best limit order and 2nd best limit order. Then we construct the variable,

$$y_t^{(n)\top} = [\Delta\tilde{p}_t^{(n)}, \Delta\tilde{s}_t^{a1(n)}, \Delta\tilde{s}_t^{a2(n)}, \Delta\tilde{s}_t^{a3(n)}, \Delta\tilde{s}_t^{b1(n)}, \Delta\tilde{s}_t^{b2(n)}, \Delta\tilde{s}_t^{b3(n)}] \tag{5.2}$$

where $\Delta\tilde{p}_t^{(n)}$ is the prepared price factor for stock n, $\Delta\tilde{s}_t^{ar(n)}$ stands for the corresponding rth level of ask size factor, whereas $\Delta\tilde{s}_t^{br(n)}$ stands for the rth level of bid size factor for stock n.

By stacking the vector $y_t^{(n)\top}$ for different N stocks together, we define the large vector $Y_t^\top$ to estimate as

$$Y_t^\top = [y_t^{(1)\top}, y_t^{(2)\top}, \dots, y_t^{(N)\top}] \tag{5.3}$$

Note that a critical assumption imposed to ensure the consistency of estimator is the observations are weakly dependence. In our setting we divide the trading period into 1-minute intervals and pre-average both $\Delta\tilde{p}_t^{(n)}$, $\Delta\tilde{s}_t^{br(n)}$ and $\Delta\tilde{s}_t^{ar(n)}$ to reduce microstructure noise over 15-min, yielding 375 observations per day.

5.3 Methodology

5.3.1 High-dimensional VAR estimation

Statistically, a high-dimensional (HD) VAR model facilitates consistent estimation and better finite-sample performance. Economically speaking, estimation results derived from a sparsity assumption help to explain the economic intuition. By incorporating the lags terms in the penalized VAR model, we aim to show the "sluggished" price adjustments caused by market/limit orders.

The standard VAR(p) model, Lütkepohl (2007) is,

$$\begin{aligned} Y_t &= A_1 Y_{t-1} + A_2 Y_{t-2} + \dots + A_p Y_{t-p} + u_t \\ &= (A_1, A_2, \dots, A_p) \left(Y_{t-1}^\top, Y_{t-2}^\top, \dots, Y_{t-p}^\top\right)^\top + u_t \end{aligned} \tag{5.4}$$

where $Y_t = (y_{1t}, y_{2t}, \dots, y_{Kt})^\top \in \mathbb{R}^K$ is a random vector, $t = 1, \dots, T$. A_i are fixed $(K \times K)$ coefficient matrices. p is the lag and $u_t = (u_{1t}, u_{2t}, \dots, u_{Kt})^\top \in \mathbb{R}^K$ is the i.i.d innovation process. In our LOB setting, dimension of $K = 7N$ with N is the number of stocks in the portfolio.

Assumption 1. Assume (5.4) satisfies that,

1. The roots of $|I_K - \sum_{j=1}^{p} A_j z^j| = 0$ lie outside unit circle.
2. u_t are i.i.d innovations;
   each element has bounded $(4+\delta)$th moment, $\delta > 0$.
3. $\|\Sigma_u\|_2 < \infty$ and $\|(A_1, A_2, \dots, A_p)\|_2 < \infty$.

In practice, the coefficients $A_1, \dots, A_p$ are unknown and have to be estimated from $\{Y_t\}_{t=1}^{T}$. Define,

$$\begin{array}{rcll} Y & = & (Y_1, Y_2, \dots, Y_T) & A = (A_1, A_2, \dots, A_p) \\ Z_t & = & (y_t, y_{t+1}, \dots, y_{t-p+1})^\top & Z = (Z_0, Z_1, \dots, Z_{T-1}) \end{array} \tag{5.5}$$

Then equation (5.4) reads,

$$Y = AZ + U \tag{5.6}$$

with $U = (u_1, u_2, \dots, u_T)$. The compact form (5.6) is equivalent to

$$\mathbf{y} = (Z^\top \otimes I_K)\beta + \mathbf{u} = \mathbf{x}\beta + \mathbf{u} \tag{5.7}$$

where the length of the parameter vector β is pK^2, the number of observations is KT.

In practice, the ration $\frac{Kp}{T}$ could be large due to high dimensionality, which deteriorates the accuracy of final estimate. Worse still, if $Kp > T$, the number of coefficients to be estimated increases quadratically in terms of the number of lags p, therefore the model cannot be identified with traditional methods such as OLS. Therefore variable selection techniques like Lasso is introduced to concentrate on a subset of non-zero parameters. For multiple time series data, especially high dimensional time series, it is preferred to use elastic net approach rather than pure Lasso to remedy potentially strong correlation among regressors. Besides, under normal assumption of error term, the upper bound of estimated error is positively correlated in $\frac{\log(K^2 p)}{T}$, part of oracle inequality. The methodologies introduced in the proceeding paragraph are of great importance in the sense that the true underlying model has a sparse representation.

The HD VAR estimates β by minimizing the objective function,

$$\arg\min_{\beta} \left(\|\mathbf{y} - \mathbf{x}\beta\|_2^2 + \alpha_{1,T}\|\beta\|_1 + \alpha_{2,T}\|\beta\|_2^2 \right) \tag{5.8}$$

which is equivalent to,

$$\underset{A_1,A_2,\ldots,A_P}{\arg\min} \sum_{t=1}^{T} \|Y_t - \sum_{j=1}^{P} A_j Y_{t-j}\|_2^2 + \alpha_{1,T} \sum_{j=1}^{P} \|vec(A_j)\|_1 + \alpha_{2,T} \sum_{j=1}^{P} \|vec(A_j)\|_2^2 \quad (5.9)$$

where A_j is the $(K \times K)$ coefficient matrices of interest. $\alpha_{1,T}$ and $\alpha_{2,T}$ are the penalty parameters. Note that the notation $\|M\|_p$ depends on whether M is a vector or a matrix. To avoid confusion, we use $vec(M)$ here to tranform the object within $\|\|_p$ into a vector.

We choose a sequence of decreasing positive numbers $\alpha_{1,T}$ and $\alpha_{2,T}$ to control the regularization. In the case of regularization parameter is large, setting it too high will throw away useful information, whereas the estimated graph is not sparse when the α_T is small. To balance the sparsity and estimation accuracy, we choose a moderately small tuning parameter using the Bayesian information criterion (BIC). In addition, we apply OLS post-model selection estimator to the first-step penalized estimator (5.8) or (5.9) to reduce shrinkage bias and ensure better model model performance.

5.3.2 Structural Analysis of High-dimensional LOB Portfolio

This section discusses the effects of uncertainty shocks in the LOB. In general, uncertainty responds to all shocks through its relation to the lags of the LOB variables as specified in the HD VAR model (5.8). Let us first consider the generalized impulse response function (GI) for the case of an arbitrary current shock.

For the multivariate case, following Koop et al. (1996) and Pesaran and Shin (1998), we assume shocks hitting only one equation at a time rather than all the shocks at time t. The effect on j-th equation of y_t of a one-standard deviation shock to perceived uncertainty are given by GI,

$$\begin{aligned} \delta_{jt} &: (\delta_{1t}, \delta_{2t}, \ldots, \delta_{Kt})^\top \sim \hat{u}^\star_{jt} e_j \qquad (5.10) \\ GI(l, \delta_{jt}, \mathcal{F}_{t-1}) &= \mathsf{E}(y_{t+l} \mid u_{jt} = \delta_{jt}, \mathcal{F}_{t-1}) - \mathsf{E}(y_{t+l} \mid \mathcal{F}_{t-1}) \end{aligned}$$

where $\hat{u}^\star_{jt}$ are independent draws with replacement from the set of residuals $\{\hat{u}_{jt}\}_{t=1}^{T}$ over the sample period, with $\{\hat{u}_{jt}\}$ is the model-implied residual of jth equation at time t. $\mathsf{E}(y_{t+1} \mid u_{jt} = \delta_{jt}, \mathcal{F}_{t-1})$ represents the expectation conditional on the history $\mathcal{F}_{t-1}$ and a fixed value of j-th shock δ_{jt} on time t at horizon l. $\mathcal{F}_{t-1}$ consists of the information used to compute the conditional expectations based on (5.4).

To measure the persistent effect of a shock on the behaviour of a series, the basic object in 5.10 is the conditional expectation. However the sparse estimation of HD VAR is nonlinear, the GI functions cannot be expressed in closed form. Therefore we use bootstrap methods to produce simulated realizations that can be used to form draws from the joint distribution of shocks. The steps for computing the conditional expectations in GI are described in Appendix 5.B.

5.3.3 Network Construction

The LN-GFEVD denoted as $\lambda_{ij,\mathcal{F}_{t-1}}(h)$ is defined by j-th shock hitting i-th variable at time t,

$$\lambda_{ij,\mathcal{F}_{t-1}}(h) = \frac{\sum_{l=0}^{h} GI(l, \delta_{jt}, \mathcal{F}_{t-1})_i^2}{\sum_{j=1}^{K} \sum_{l=0}^{h} GI(l, \delta_{jt}, \mathcal{F}_{t-1})_i^2}, \quad i, j = 1, \ldots, K \tag{5.11}$$

where h is the horizon, $\mathcal{F}_{t-1}$ refers to the history. Therefore $\lambda_{ij,\mathcal{F}_{t-1}}(h) \in [0, 1]$, measuring the relative contribution of a shock δ_{jt} to the j-th equation in relation to the total impact of all K shocks on the i-th variable in y_t after h periods. Compared to traditional GFEVD, LN-GFEVD has the attractive property that the proportions of the impact accounted for by innovations in each variable sum to unity. The LN-GFEVD is thus economic interpretable.

Many literature characterizes connectedness of the variables in the VAR systerm, for instance, Diebold and Yılmaz (2014) and Demirer et al. (2017) proposed connectedness measures built from GFEVD for both univariate and multivariate cases. However, to our knowledge, the combination of bootstrap-based GI analysis and network construction seems to be new to the literature: Upon the HD VAR estimation of (5.8) and (5.9), we use the sparsity concept that filters out less relevant variables. Instead of transforming into a MA process, which is often done in fixed dimensional cases, we apply a bootstrap-based method to compute $\lambda_{ij,\mathcal{F}_{t-1}}(h)$, then naturally produce the population connectedness, see Table 5.3. By this way, the connectedness table can be constructed for both linear and nonlinear models. Besides, the bootstrapped LN-GFEVD relies neither on the ordering of the variables nor on the distribution of the innovations. At the same time, a parsimonious algorithm without MA transformation can help to improve the accuracy of final connectedness estimator.

The details for computation steps can be found in Appendix 5.B. In particular, the numerical techniques for conditional mean forecast from nonlinear models for more than one period ahead are implemented in this paper, we use bootstrap to calculate $GI(l, \delta_{jt}, \mathcal{F}_{t-1})$, see more details in Teräsvirta et al. (2010).

	x_1	x_2	$\dots$	x_K	From others
x_1	$\lambda^b_{11}(h)$	$\lambda^b_{12}(h)$	$\dots$	$\lambda^b_{1K}(h)$	$\sum_{j=1}^K \lambda^b_{1j}(h), j \neq 1$
x_2	$\lambda^b_{21}(h)$	$\lambda^b_{22}(h)$	$\dots$	$\lambda^b_{2K}(h)$	$\sum_{j=1}^K \lambda^b_{2j}(h), j \neq 2$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_K	$\lambda^b_{K1}(h)$	$\lambda^b_{K2}(h)$	$\dots$	$\lambda^b_{KK}(h)$	$\sum_{j=1}^K \lambda^b_{Kj}(h), j \neq K$
To others	$\sum_{i=1}^K \lambda^b_{i1}(h)$ $i \neq 1$	$\sum_{i=1}^K \lambda^b_{i2}(h)$ $i \neq 2$	$\dots$	$\sum_{i=1}^K \lambda^b_{iK}(h)$ $i \neq K$	$\frac{1}{K}\sum_{i=1,j=1}^K \lambda^b_{ij}(h)$ $i \neq j$

Table 5.3. Connectedness table of interest, estimated by bootstrap-based methods.

We then have the directional connectedness "from" and "to" associated with the forecast error variation $\lambda^b_{ij}(h)$ for a specific order book across various stock when the arising shocks transmit from one stock to the others. These two connectedness estimators can be obtained by adding up the row or column elements. Hence the pairwise directional connectedness from j to i can be written as,

$$C_{i \leftarrow j} = \lambda^b_{ij}(h) \tag{5.12}$$

Furthermore, the total directional connectedness "from" $C_{i\leftarrow \cdot}$ (others to i) given by

$$C_{i\leftarrow \bullet} = \sum_{j=1}^{K} \lambda^b_{ij}(h), i \neq j \tag{5.13}$$

equals to unity based on (5.11), and the total directional connectedness "to" $C_{\cdot \leftarrow j}$ (j to others) is defined as

$$C_{\bullet \leftarrow j} = \sum_{i=1}^{K} \lambda^b_{ij}(h), i \neq j \tag{5.14}$$

The corresponding net total directional connectedness

$$C_i = C_{to,i} - C_{from,i} = C_{\bullet \leftarrow i} - C_{i \leftarrow \bullet} \tag{5.15}$$

measures the direction and magnitude of the net spillover impacts.

5.4 Network Analysis

Upon the estimates of the sparse HD VAR model, we calculate the bootstrapped LN-GFEVD and corresponding connectedness at horizon $h = 30$ for every trading day.

5.4.1 Individual Stock Network

Pairwise Connectedness of Stocks

Let us first focus on the individual stock network to understand how the impact of a shock originating in one stock can be transmitted and amplified to the other stocks.

Basically a network can be considered as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consisting of two core items: nodes (or vertexs) $\mathcal{V}$ and edges $\mathcal{E}$. Nodes are the entities we are evaluating and edges are the connections between them. Here we first consider the cross-stock network $\mathcal{G}_p = (\mathcal{V}_p, \mathcal{E}_p)$ with only price factors $p^{(n)}$,

$$\mathcal{V}_p = p^{(n)}, \quad n = 1, \ldots, N \quad \text{and} \quad \mathcal{E}_p = C_{i \leftarrow j}, \quad i, j \in \mathcal{V}_p \tag{5.16}$$

We model each trading day as a separate network and extract the pairwise connectedness estimate for each stock. To understand the behavior of networks, there are various approaches for evaluating the node importance. We employ the centrality measures proposed by Freeman (1978) to evaluate the relative importance of nine stocks,

- degree centrality $deg(\mathcal{V})$: refers to the number of edges attached to one node. This is simplest measure of node connectivity, but it is can be interpreted as a form of popularity. We use “out-degree” centrality $outdeg(\mathcal{V})$, i.e. the number of ties that the node directs to others to measure the impact of “to”-connectedness, and “in-degree” centrality $indeg(\mathcal{V})$ (number of inbound links) to measure the impact of “from”-connectedness.
- closeness centrality $Clos(\mathcal{V})$: is defined as the inverse of the sum of its distances to all other nodes, it scores each node based on their closeness to all other nodes within the network. Thus we are able to identify the nodes who are best placed to influence the entire network most quickly. The more central a node is, the closer it is to all other nodes. This centrality measure will be useful to distinguish influencers in the network.
- betweenness centrality $Bet(\mathcal{V})$: quantifies the number of times a node lies on the shortest path between other nodes. Nodes that have a high probability to occur on a randomly chosen shortest path between two randomly chosen vertices have a high betweenness. This centrality measure is helpful to decide which nodes act as “bridges” between nodes in a network, and can potentially influence the spread of information through the network.

To better grasp the results, given the large amount of estimation results, we will use the summary results in tables throughout the paper. Table 5.4 provides the summary of the corresponding centrality measures. Citigroup, AT&T and Johnson&Johnson are central in the network, in the sense that nodes with higher "out-degree" play the role of choice maker. Meanwhile JNJ is a choice receiver with high "in-degree" value of 3.57, slightly smaller than 3.88 of Microsoft. JP Morgan and IBM are the nodes who are best placed to influence the entire network most quickly, with IBM acts as "bridge" between nodes at the same time. The above conventional centrality measures are helpful to understand the evolution of the pairwise network, but we cannot accurately classify the most important nodes demonstrating the high centrality values with above results. Even though each measure works well for probing certain phenomena, it fails to capture the node's spreading potential, e.g. Johnson&Johnson.

	MSFT	T	IBM	JNJ	PFE	MRK	JPM	WFC	C
$Q_{outdeg(\mathcal{V}_p)}(0.25)$	2.00	3.00	1.25	2.00	2.00	1.00	1.00	2.00	2.00
$Q_{outdeg(\mathcal{V}_p)}(0.75)$	4.00	4.00	4.00	5.00	4.00	4.00	4.00	4.00	5.00
$\mu_{outdeg(\mathcal{V}_p)}$	3.26	3.33	2.52	3.33	3.07	2.81	2.83	2.50	3.40
$Q_{indeg(\mathcal{V}_p)}(0.25)$	2.25	1.00	2.00	2.00	1.00	0.00	1.00	0.00	2.00
$Q_{indeg(\mathcal{V}_p)}(0.75)$	6.00	4.75	5.75	5.00	3.00	3.75	5.00	4.75	5.00
$\mu_{indeg(\mathcal{V}_p)}$	3.88	2.86	3.43	3.57	2.38	1.69	3.21	2.60	3.45
$Q_{Clos(\mathcal{V}_p)}(0.25)$	12.56	20.08	19.72	15.26	18.94	14.27	16.93	20.08	15.33
$Q_{Clos(\mathcal{V}_p)}(0.75)$	254.51	228.13	265.74	257.15	242.92	211.66	283.35	237.15	237.09
$\mu_{Clos(\mathcal{V}_p)}$	167.70	163.99	173.45	159.09	159.43	154.56	175.89	171.95	157.81
$Q_{Bet(\mathcal{V}_p)}(0.25)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$Q_{Bet(\mathcal{V}_p)}(0.75)$	10.50	14.75	14.75	6.50	5.00	4.75	10.00	9.00	10.00
$\mu_{Bet(\mathcal{V}_p)}$	6.00	7.55	7.98	4.33	4.43	3.02	5.98	5.24	6.07

Table 5.4. Summary of different centrality measures for $\mathcal{G}_p$ from 06.2016 to 07.2016. $Q_{\cdot}(\alpha)$ is the quantile function, $\mu_{\cdot}$ is the mean.

Including Order Flows

We now investigate how the network is affected by the presence of liquidity effects, i.e. by including the order volumes in the book.

We take the first trading day after Brexit as an example. In accordance with the discussion in section 5.3.3, we depict the estimated full sample directional connectedness Table 5.3 in left panel of Figure 5.3. Directed connectedness are drawn as directed lines connecting two nodes. The price factor and size factors that belong to the same company appear in the same colour, the width of edges between two nodes represents the connectedness. The full sample network plot reveals that the stocks with LOB factors are massively connected, it is quite informative about the total directional connectedness of each factor. However, it is not easy to decipher all pairwise connectedness. On the right panel, each stock is a node in the network, links between nodes represent the overall "from"' and "to" impacts on the system, i.e., aggregating the connectedness measure of both price and size factors for each stock. The respective links of Citigroup and JP Morgan and Wells Fargo reveal that they are the stocks that generated highest "to"-connectedness, whereas the other six stocks are mainly risk receiver.

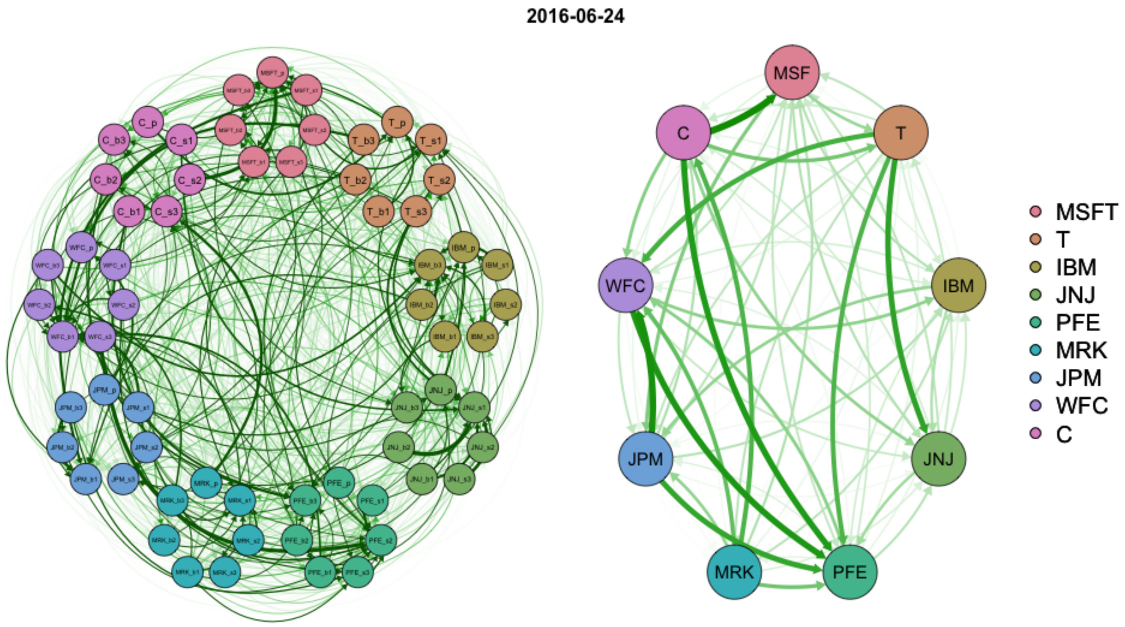


Figure 5.3. Left panel: the full sample network plot. Right panel: the aggregated network plot of nine stocks, on 24.06.2016

To formalize the analysis we construct the network based on (5.16), the aggregated individual stock network is given by $\mathcal{G}_g = (\mathcal{V}_g, \mathcal{E}_g)$ consisting of,

$$\mathcal{V}_g = v_g^{(n)} \tag{5.17}$$

$$v_g^{(n)} = p^{(n)} + \sum_r bs_r^{(n)} + \sum_r as_r^{(n)}, \quad n = 1, \dots, N \tag{5.18}$$

$$\mathcal{E}_g = C_{i \leftarrow j}, \quad i, j \in \mathcal{V}_g \tag{5.19}$$

where $as_r^{(n)}$ and $bs_r^{(n)}$ are the r-th level ask/bid size factors for stock n. By including the size factors from LOB, we are able to investigate how the network is affected by the presence of liquidity effects. For a network with smaller number of nodes, it is easy and appealing to identify the characteristics and patterns between individual stock.

Table 5.5 tabulates the centrality measures based on aggregated nine stock network, which produces different results comparing to those obtained for pairwise stock network in 5.4.1. The primary reason is that these conventional centrality measures are rarely accurate when the majority of nodes are not highly influential in the network. Each centrality measure assess the node's importance based mostly on the path lengths and distances. The impacts caused by the less important nodes may be neglected, this will potentially cause inaccuracy and thus result in the poor performance. Therefore we use net total directional connectedness proposed in (5.15) as a refined centrality measure to capture the most influential spread in the following full sample network analysis.

	MSFT	T	IBM	JNJ	PFE	MRK	JPM	WFC	C
$Q_{outdeg(\mathcal{V}_g)}(0.25)$	1.00	1.00	1.00	1.00	0.00	0.25	0.00	0.00	0.00
$Q_{outdeg(\mathcal{V}_g)}(0.75)$	114.50	131.00	115.75	111.00	110.00	103.75	105.75	98.00	105.25
$\mu_{outdeg(\mathcal{V}_g)}$	128.83	147.02	132.71	129.76	127.95	125.48	120.31	113.50	123.00
$Q_{indeg(\mathcal{V}_g)}(0.25)$	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$Q_{indeg(\mathcal{V}_g)}(0.75)$	100.50	89.75	97.00	100.75	96.50	90.50	111.25	108.00	98.00
$\mu_{indeg(\mathcal{V}_g)}$	136.29	122.50	121.24	118.31	121.88	121.00	136.79	133.98	136.60
$Q_{Clos(\mathcal{V}_g)}(0.25)$	0.02	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.02
$Q_{Clos(\mathcal{V}_g)}(0.75)$	0.08	0.08	0.07	0.08	0.09	0.08	0.08	0.08	0.09
$\mu_{Clos(\mathcal{V}_g)}$	0.13	0.13	0.12	0.13	0.13	0.13	0.13	0.13	0.13
$Q_{Bet(\mathcal{V}_g)}(0.25)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$Q_{Bet(\mathcal{V}_g)}(0.75)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\mu_{Bet(\mathcal{V}_g)}$	4.17	3.26	2.36	2.07	2.79	2.19	3.12	3.57	3.69

Table 5.5. Summary of different centrality measures for $\mathcal{G}_g$ from 06.2016 to 07.2016. $Q_{\cdot}(\alpha)$ is the quantile function, $\mu_{\cdot}$ is the mean.

Specifically, the element in the connectedness table measures the total impact of all K shocks on the i-th variable, and these contributions sum to unity, which suggests the row sum of the pairwise connectedness produces one unit of "from"-connectedness for each factor, therefore the "net"-connectedness C_i is associated with "to"-connectedness and measures the share of volatility shocks to other. To understand the dynamic behavior of the risk transmission in the system, Table 5.6 reports the net spillover effects for each stock using the quantile functions,

$$\begin{aligned} C_i &= C_{\bullet\leftarrow i} - 2r - 1 = \sum_j C_{j\leftarrow i} - 2r - 1, \quad i, j \in \mathcal{V}_g \\ Q_{C_i}(\alpha) &= F^{-1}(\alpha) = \inf\{C_i : F(C_i) \geqslant \alpha\} \end{aligned} \tag{5.20}$$

In the table, JP Morgan is the stock with the highest "net" connectedness to others, with mean value of 0.97 over the sample period, followed by Citigroup 0.35, Wells Fargo 0.34, Microsoft 0.17, Pfizer 0.14. The "net" total connectedness of the left four stocks are all negative. As an evident result one see that that the JP Morgan is most influential in the network, while the technology companies like IBM and AT&T are main risk receivers in the aggregated system. Even though the magnitude of fi-

	MSFT	T	IBM	JNJ	PFE	MRK	JPM	WFC	C
$Q_{C_i}(0.05)$	-3.19	-3.51	-3.72	-3.79	-2.84	-3.38	-2.57	-2.80	-3.35
$Q_{C_i}(0.15)$	-2.15	-3.01	-3.24	-2.99	-2.16	-2.62	-1.97	-2.30	-2.61
$Q_{C_i}(0.50)$	0.17	-0.70	-0.91	-0.50	0.14	-0.71	0.97	0.34	0.35
$Q_{C_i}(0.85)$	2.01	1.47	1.31	2.27	1.78	2.47	3.68	3.21	3.04
$Q_{C_i}(0.95)$	3.84	2.54	2.99	2.70	3.81	3.28	5.11	4.63	4.74

Table 5.6. The net spillover of nine-stock aggregation from 06.2016-07.2016

nancial stock estimates differs to some extent, their "net"-connectedness are larger than the other in most cases. This suggests that financial companies are dominant stocks driving the networks over time. We conclude that the sign and magnitude of "net"-connectedness provide different information regarding the role for each stock in the network. The aggregated individual stock network is a better measure of how central a stock is within the network since it takes into consideration the trading volumes.

Total Connectedness and Volatility

We now turn our focus on time-varying pattern of the aggregated individual stock network in comparison with daily volatility estimates using the full sample high-frequency observations. Estimating volatility in this context is important as they are commonly known as proxies of market fear, a high degree of volatility is likely to correspond to increasing market risk and represent the market consensus on the expected future uncertainty.

Inspired by a voluminous literature such as Andersen et al. (2000), Andersen and Bollerslev (1998), Andersen et al. (2001) and Barndorff-Nielsen and Shephard (2001), the realized volatility (RV) is illustrated as measure of daily volatility in high-frequency setting. In literature, several main approaches to improve the realized volatility (RV) estimator include the preaveraging estimator of Jacod et al. (2009), the realized kernel estimator of Barndorff-Nielsen et al. (2008), the two scales estimator of Zhang et al. (2005) and multiscale estimator of Zhang et al. (2006) and Zhang (2011). Here we compute the two-scale realized variance (TS-RV) proposed by Zhang et al. (2005) as a robust estimator of the RV. The TS-RV estimator computes a subsampled RV on one slower time scale and then combine with another subsample RV calculated on a faster time scale to correct for microstructure noise.

Figure 5.4 compares the total net connectedness with estimated daily volatility,

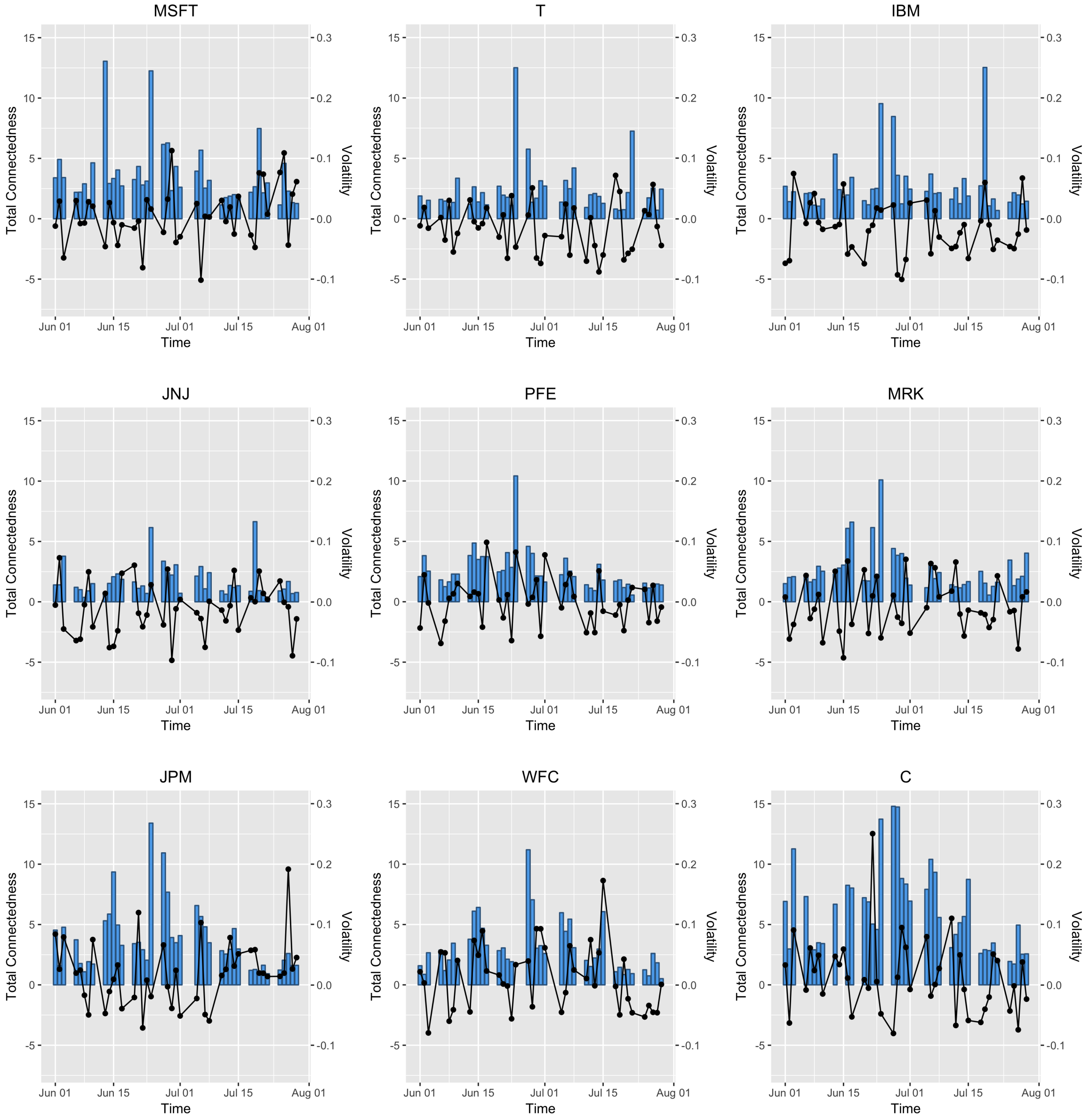


Figure 5.4. The time varying total net connectedness and volatility measure, 06-07.2017

where the dotted lines illustrate the total connectedness estimates of (5.20), and the barplots indicate the TS-RV estimates. Visual inspection of the time series plots suggests, for all stocks, a rising volatility phase since the beginning of June, with the peak volatility observed around 24th of June, after that volatility decreases given the selloff in stocks following the Brexit vote followed by a rebound to record highs. The findings are consistent with the results of net connectedness measures, where a very small value of C_i is usually observed near Brexit: in other words the stocks are less connected when high market volatility occurs. In addition, we observe another peaks in volatility appear around 18th of July 2016 for three technology stocks, when Turkish shares closed down by 7.1% following the attempted coup in Turkey on 17th of July 2016. Then the volatility level come back in as the market fear caused by after coup attempt in Turkey is resolved. While important events play out, investors are likely to join a selloff as geopolitical risk is always important for decision-making in financial market. Since the peaks of volatility are generally correspond to a very low net connectedness value, this can be a signal for market investors because a peak in volatility is followed by a market rally in most cases.

5.4.2 Limit Order Book Network

Asymmetric Market Sell/Buy Pressure

Besides the purpose of studying the impacts between individual stocks, the information contained in the LOB is very valuable. Limit orders are stored in the LOB and are executed in sequence according to price priority, large trading quantities may cause a price drop or rise. The intuition behind a typical mechanism resulting in mid-price movement can be illustrated in combination with Figure 5.1. If there is an arrival of a market order that is sufficiently large to match all of the best bids, then the limit order will be updated with a lower best bid price.

Figure 5.5 shows the graphical display of the networks consisting of price factors, ask size factors and bid size factors, with the connectedness $C_{i\leftarrow j}$ color-coded by the type of factors that is causing the relationship, i.e., the factor j which has an impact on the others. Blue indicates the ask size factors, red indicates the bid size factors, and grey indicates the price factors. The upper left panel of Figure 5.5 depicts the full-sample connectedness on 22.06.2016, which is hard to decipher important pairwise connectedness. Therefore we decompose the full-sample connectedness into two parts, the price&ask size connectedness graph and price&bid size connectedness graph as shown in colored circles on the right panel. It shows how the LOB network changed during Brexit announcement, we typically observe changes in the behavior of bid size factors. The price&bid size factor network is less connected on 23.06.2016, while the price&ask size factor network is slightly tightly connected on the same day.

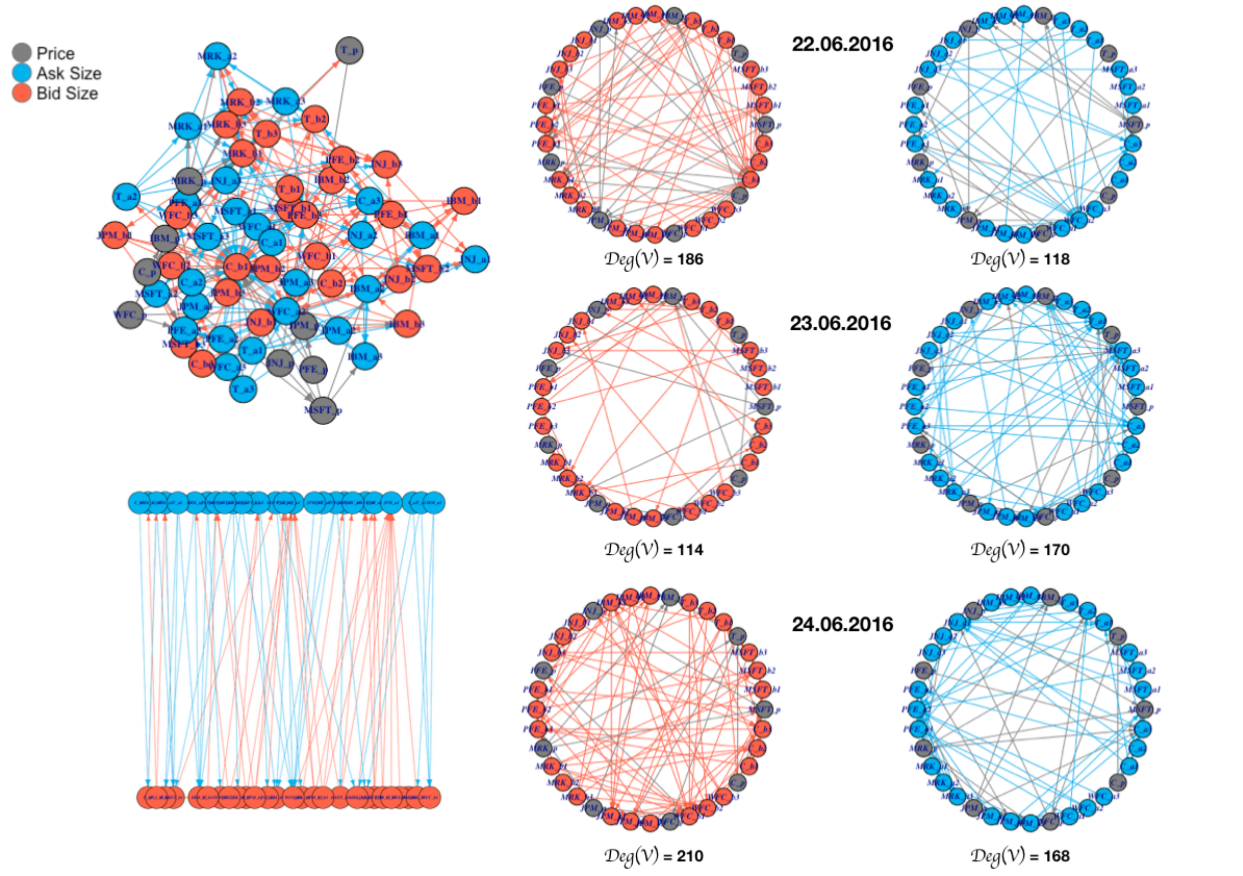

Price
Ask Size
Bid Size
22.06.2016
$\mathcal{D}eg(\mathcal{V})$ = 186
$\mathcal{D}eg(\mathcal{V})$ = 118
23.06.2016
$\mathcal{D}eg(\mathcal{V})$ = 114
$\mathcal{D}eg(\mathcal{V})$ = 170
24.06.2016
$\mathcal{D}eg(\mathcal{V})$ = 210
$\mathcal{D}eg(\mathcal{V})$ = 168

Figure 5.5. Plots of LOB networks from 22.06.2016-24.06.2016

This result could indicate that, when there is a risk caused by political uncertainty, the buying pressure is much weaker and selling pressure slightly stronger.

The impacts on returns respond to ask and bid limit orders are not symmetric. Recent studies have showed that limit orders and cancelations, not just trades, have a tangible effect on prices, see Hautsch and Huang (2012) and Eisler et al. (2012). Building on these ideas, we construct a graph $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s)$ to study the asymmetric impact from aggregated size factors to price factors,

$$\mathcal{V}_s = \left(p^{(n)}, \sum_n bs_r^{(n)}, \sum_n as_r^{(n)}\right) \quad n = 1, \dots, N \tag{5.21}$$

$$\mathcal{E}_s = C_{i \leftarrow j} \quad i \in \{p^{(n)}\}, \quad j \in \left\{\sum_n bs_r^{(n)}, \sum_n as_r^{(n)}\right\} \tag{5.22}$$

	MSFT	T	IBM	JNJ	PFE	MRK	JPM	WFC	C
$QC_{p^{(n)} \leftarrow \sum as1}(0.25)$	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.00
$QC_{p^{(n)} \leftarrow \sum as1}(0.75)$	0.31	0.41	0.39	0.26	0.42	0.34	0.72	0.45	0.43
$\mu C_{p^{(n)} \leftarrow \sum as1}$	0.44	0.43	0.45	0.34	0.45	0.68	0.86	0.53	0.83
$QC_{p^{(n)} \leftarrow \sum as2}(0.25)$	0.02	0.02	0.00	0.01	0.01	0.02	0.01	0.01	0.01
$QC_{p^{(n)} \leftarrow \sum as2}(0.75)$	0.61	0.50	0.58	0.43	0.26	1.00	0.30	0.24	0.40
$\mu C_{p^{(n)} \leftarrow \sum as2}$	0.53	0.41	0.59	0.48	0.60	0.65	0.33	0.50	0.64
$QC_{p^{(n)} \leftarrow \sum as3}(0.25)$	0.02	0.02	0.00	0.01	0.01	0.02	0.03	0.01	0.02
$QC_{p^{(n)} \leftarrow \sum as3}(0.75)$	0.71	0.52	0.54	0.34	0.37	0.50	0.88	0.45	0.46
$\mu C_{p^{(n)} \leftarrow \sum as3}$	0.90	0.69	0.54	0.33	0.58	0.89	1.03	0.54	0.39
$QC_{p^{(n)} \leftarrow \sum bs1}(0.25)$	0.01	0.02	0.01	0.01	0.00	0.01	0.01	0.00	0.01
$QC_{p^{(n)} \leftarrow \sum bs1}(0.75)$	0.11	0.34	0.21	0.20	0.23	0.28	0.26	0.66	0.42
$\mu C_{p^{(n)} \leftarrow \sum bs1}$	0.12	0.44	0.47	0.40	0.32	0.41	0.46	0.61	0.40
$QC_{p^{(n)} \leftarrow \sum bs2}(0.25)$	0.00	0.02	0.00	0.00	0.01	0.01	0.01	0.02	0.01
$QC_{p^{(n)} \leftarrow \sum bs2}(0.75)$	0.26	0.29	0.31	0.11	0.46	0.27	0.33	0.16	0.26
$\mu C_{p^{(n)} \leftarrow \sum bs2}$	0.50	0.35	0.38	0.20	0.37	0.22	0.48	0.20	0.28
$QC_{p^{(n)} \leftarrow \sum bs3}(0.25)$	0.01	0.02	0.00	0.01	0.00	0.01	0.01	0.03	0.01
$QC_{p^{(n)} \leftarrow \sum bs3}(0.75)$	0.43	0.40	0.15	0.20	0.32	0.21	0.49	0.71	0.45
$\mu C_{p^{(n)} \leftarrow \sum bs3}$	0.63	0.37	0.24	0.39	0.51	0.38	0.43	0.73	0.47

Table 5.7. Summary of the aggregated impacts from size factors to the stock price factor from 06.2016-07.2017. $Q.(\alpha)$ is the quantile function, $\mu.$ is the mean.

In Table 5.7 we provide the summary of $\mathcal{E}_s$ in (5.22), i.e. impacts from aggregated size factors to the stock price factor. The higher are the values in this table, the stronger are the stocks affect by trading activities over time. We notice that JP Morgan on average is more likely to be affected by ask side trading activity, while Wells Fargo is most sensitive to the bid side trading activity. In addition, both best bid and ask limit orders (i.e. 2nd level of ask/bid size) exhibit opposite results, with JP Morgan and Wells Fargo are less likely to be affected by best ask and bid limit order respectively. We find this result very interesting, because it brings into question how best limit orders are correlated with the order flow preceding their arrival and therefore have very little impacts on the price. This may be explained by the assumption that both market and limit orders tend to drive prices, while prices tend to impact best limit orders and their cancellations in the book. We conclude that the financial stocks are size-dominated stocks, their price patterns are highly related to the market trading activity. When selling pressure increases, the Healthcare stocks are more stable. While the technology stocks appear to be more stable for buying pressure.

It follows that the depth of the book at which limit orders are submitted is driving the price. Accordingly, we calculate the impacts from the aggregated ask/bid size factors to the aggregated price factors given by,

$$C_{\sum_N p \leftarrow \sum_N s_r^{(n)}} = \sum_{i=1}^{N} C_{i \leftarrow j} \tag{5.23}$$

$$i \in \{p^{(n)}\}, \quad j \in \left\{ \sum_n bs_r^{(n)}, \sum_n as_r^{(n)} \right\} \tag{5.24}$$

Table 5.8 compares the aggregated impacts for six types of size factors in our study. The impacts on return (aggregated price factors) respond to incoming ask and bid market/limit orders are not symmetric. In general, the impacts from ask orders are larger than the bid orders, ranging from the lowest value of 0.30 for aggregated impacts of bs_2 to the highest value of 0.59 for as_3 on average. Please note that this results are consistent with the results of Table 5.7, indicating that we can observe stronger impacts on prices caused by market sell pressure.

	$Q_C(0.25)$	$Q_C(0.50)$	$Q_C(0.75)$	μ_C
$C_{\sum_N p \leftarrow \sum_N as_1^{(n)}}$	0.12	0.27	0.67	0.50
$C_{\sum_N p \leftarrow \sum_N as_2^{(n)}}$	0.19	0.30	0.55	0.47
$C_{\sum_N p \leftarrow \sum_N as_3^{(n)}}$	0.17	0.35	0.83	0.59
$C_{\sum_N p \leftarrow \sum_N bs_1^{(n)}}$	0.09	0.18	0.39	0.36
$C_{\sum_N p \leftarrow \sum_N bs_2^{(n)}}$	0.08	0.16	0.41	0.30
$C_{\sum_N p \leftarrow \sum_N bs_3^{(n)}}$	0.13	0.29	0.63	0.42

Table 5.8. Summary of the impacts from aggregated size factors to the aggregated price factor from 06.2016-07.2017. $Q_{\cdot}(\alpha)$ is the quantile function, $\mu_{\cdot}$ is the mean.

More precisely, let μ_1 be the mean of the overall impacts from selling orders over the sample period ($T = 42$), and μ_2 the corresponding mean of the overall impacts from buying orders, i.e.,

$$\mu_1 = \frac{1}{3T} \left(C_{t, \sum_N p \leftarrow \sum_N as_1^{(n)}} + C_{t, \sum_N p \leftarrow \sum_N as_2^{(n)}} + C_{t, \sum_N p \leftarrow \sum_N as_3^{(n)}} \right) \tag{5.25}$$

$$\mu_2 = \frac{1}{3T} \left(C_{t, \sum_N p \leftarrow \sum_N bs_1^{(n)}} + C_{t, \sum_N p \leftarrow \sum_N bs_2^{(n)}} + C_{t, \sum_N p \leftarrow \sum_N bs_3^{(n)}} \right) \tag{5.26}$$

therefore the hypothesis of interest can be expressed as,

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = 0 \\ H_a &: \mu_1 - \mu_2 > 0 \end{aligned}$$

Table 5.9 suggests that both the pooled t-test and the Welsh t-test give roughly the same results. Since the p-value is very low, we reject the null hypothesis, indicating that there is strong evidence of a significant larger impact from selling orders in the market.

	t-statistics	p-value
Pooled t-test	2.7557	0.003144
Welsh t-test	2.7557	0.003168

Table 5.9. Comparison of two hypothesis tests, when assuming/not assuming equal standard deviation.

Own-price and Cross-price Market Impact

The discussion in section 5.4.2 concludes that the impacts on return respond to different level of depth of the book are widely asymmetric. In this subsection we provide further empirical evidence of own-price and cross-price market impact at the level on the individual stock. First, we analyze the market impacts of their own trades for each stock, and then we undertake a detailed analysis of the impact of trades in one stock on the prices of other stocks.

At first, we consider own-price market impact for different levels of depth of the book for the selected stocks, i.e. the own-price market impacts are caused by their own order flows. The results are presented in Table 5.10. Based on the averaged connectedness over two months, JP Morgan receives highest market impact from its own ask orders, especially when the orders are placed in the market order or the 2nd best limit order. Even though Wells Fargo and Microsoft are the two stocks receiving highest market impacts from their own bid trades, the market impacts from their ask trades are high as well. In addition, the Johnson&Johnson responds weakly to both ask orders and bid orders.

In contrast to (5.21), we measure the cross-price market impacts by adding up the impacts from all ask/bid orders for each stock. The graph we construct is denoted as $\mathcal{G}_{cross} = (\mathcal{V}_c, \mathcal{E}_c)$, with cross-stock market impacts from the aggregated size factors

	MSFT	T	IBM	JNJ	PFE	MRK	JPM	WFC	C
$\mu C_{p^{(n)} \leftarrow as1^{(n)}}$	0.47	0.90	1.28	0.06	0.31	2.95	3.58	2.47	0.26
$\mu C_{p^{(n)} \leftarrow as2^{(n)}}$	0.34	0.31	0.47	0.13	0.07	1.42	0.38	2.17	1.43
$\mu C_{p^{(n)} \leftarrow as3^{(n)}}$	2.57	0.26	1.30	0.25	1.69	0.90	5.26	0.74	0.32
$\sum \mu C_{p^{(n)} \leftarrow as^{(n)}}$	3.38	1.47	3.05	0.44	2.07	5.27	9.22	5.38	2.01
$\mu C_{p^{(n)} \leftarrow bs1^{(n)}}$	0.09	0.59	0.14	0.05	0.60	1.14	1.12	2.47	0.70
$\mu C_{p^{(n)} \leftarrow bs2^{(n)}}$	1.35	0.18	0.29	0.05	0.43	0.08	0.83	0.89	0.42
$\mu C_{p^{(n)} \leftarrow bs3^{(n)}}$	2.58	0.07	0.23	1.20	0.11	2.16	1.37	2.14	1.42
$\sum \mu C_{p^{(n)} \leftarrow bs^{(n)}}$	4.02	0.84	0.66	1.30	1.14	3.38	3.32	5.50	2.54

Table 5.10. The mean of own-price market impacts caused by market orders $\{as1^{(n)}, bs1^{(n)}\}$, best limit orders $\{as2^{(n)}, bs2^{(n)}\}$ and 2nd best limit orders $\{as3^{(n)}, bs3^{(n)}\}$ for each stock n from 06.2016-07.2017. All numbers are multiplied by 100. $\mu_{\cdot}$ is the mean.

to the price factor given by,

$$\mathcal{V}_c = \left(p^{(m)}, \sum_r bs_r^{(n)}, \sum_r as_r^{(n)} \right) \tag{5.27}$$

$$\mathcal{E}_c = C_{i \leftarrow j} \quad i \in \{p^{(m)}\}, \quad j \in \left\{ \sum_r bs_r^{(n)}, \sum_r as_r^{(n)} \right\} \tag{5.28}$$

$$m, n \in \{1, \dots, N\} \quad r = 1, 2, 3 \quad m \neq n \tag{5.29}$$

When $j \in \left\{ \sum_r as_r^{(n)} \right\}$ in (5.29), we compare the cross-price market impacts of ask trades for each stock in Table 5.11. Obviously, the diagonal elements measuring the market impacts of their own trades are the same as $\sum \mu C_{p^{(n)} \leftarrow as^{(n)}}$ summerised in Table 5.10. We observe three large values on the diagonal, indicating that JP Morgan, Merck and Wells Fargo have higher own-price market impacts than cross-price market impacts. Furthermore, JP Morgan is the stock with the highest cross-price market impact to Microsoft and Citigroup. IBM receives stronger cross-price market impact from Wells Fargo and Citigroup. The price of Pfizer is more sensitive to the ask order flows of Merck and Wells Fargo. Therefore we conclude that the stock price can be affected not only by their own ask order flows, but also by the ask order flows of financial stocks.

We proceed with the summary of the market impacts of bid trades for each stock when $j \in \left\{ \sum_r bs_r^{(n)} \right\}$. Table 5.12 reports the results. The table reveals that financial stocks have stronger cross-price market impacts compared with healthcare and tech-

	MSFT	T	IBM	JNJ	PFE	MRK	JPM	WFC	C
$\mu C_{p(MSFT)\leftarrow\sum_r as_r^{(n)}}$	3.38	0.68	1.82	0.65	1.76	0.73	5.46	0.54	3.66
$\mu C_{p(T)\leftarrow\sum_r as_r^{(n)}}$	3.08	1.47	1.00	0.62	1.86	2.58	1.08	1.29	2.38
$\mu C_{p(IBM)\leftarrow\sum_r as_r^{(n)}}$	1.52	0.38	3.06	1.33	0.91	1.57	1.41	3.58	2.05
$\mu C_{p(JNJ)\leftarrow\sum_r as_r^{(n)}}$	1.69	0.62	1.04	0.45	1.47	1.05	1.37	0.31	3.49
$\mu C_{p(PFE)\leftarrow\sum_r as_r^{(n)}}$	1.07	0.96	0.44	0.13	2.06	4.83	1.86	2.89	2.12
$\mu C_{p(MRK)\leftarrow\sum_r as_r^{(n)}}$	3.18	1.17	0.43	0.83	2.44	5.27	4.15	2.25	2.57
$\mu C_{p(JPM)\leftarrow\sum_r as_r^{(n)}}$	2.09	1.10	1.81	0.72	2.68	1.13	9.22	1.34	2.16
$\mu C_{p(WFC)\leftarrow\sum_r as_r^{(n)}}$	1.22	2.38	1.70	0.55	1.93	1.22	0.79	5.37	0.60
$\mu C_{p(C)\leftarrow\sum_r as_r^{(n)}}$	2.55	1.11	2.37	0.84	2.57	1.33	4.51	1.23	2.01

Table 5.11. The mean of the market impacts caused by ask orders of stock m for each stock n. All numbers are multiplied by 100. $\mu.$ is the mean.

	MSFT	T	IBM	JNJ	PFE	MRK	JPM	WFC	C
$\mu C_{p(MSFT)\leftarrow\sum_r bs_r^{(n)}}$	4.02	2.26	0.62	0.53	0.59	1.67	0.41	0.89	1.61
$\mu C_{p(T)\leftarrow\sum_r bs_r^{(n)}}$	1.36	0.84	0.22	1.04	1.41	3.67	0.92	1.10	1.03
$\mu C_{p(IBM)\leftarrow\sum_r bs_r^{(n)}}$	0.79	1.29	0.66	0.13	0.58	0.97	3.47	1.85	1.15
$\mu C_{p(JNJ)\leftarrow\sum_r bs_r^{(n)}}$	0.63	0.85	0.30	1.30	0.86	0.99	0.50	1.90	2.59
$\mu C_{p(PFE)\leftarrow\sum_r bs_r^{(n)}}$	2.12	0.36	1.10	0.19	1.13	0.37	1.43	4.08	1.23
$\mu C_{p(MRK)\leftarrow\sum_r bs_r^{(n)}}$	0.72	0.49	0.25	0.25	1.35	3.37	1.59	0.84	1.27
$\mu C_{p(JPM)\leftarrow\sum_r bs_r^{(n)}}$	1.66	0.47	1.25	0.97	1.39	0.59	3.32	1.87	2.16
$\mu C_{p(WFC)\leftarrow\sum_r bs_r^{(n)}}$	1.99	1.29	0.30	0.73	1.37	0.83	1.75	5.50	1.67
$\mu C_{p(C)\leftarrow\sum_r bs_r^{(n)}}$	1.02	1.80	0.42	1.24	0.84	1.08	1.41	1.12	2.54

Table 5.12. The mean of the market impacts caused by bid orders of stock m for each stock n. All numbers are multiplied by 100. $\mu.$ is the mean.

nology stocks. For example, the bid trades of Citigroup and Well Fargo have strong cross-price market impact on Johnson & Johnson, IBM receives stronger cross-price market impact from the bid order flows of JP Morgan and Wells Fargo.

So far we analyze the individual stock network with and without the order flows in the book, the network study enables us to investigate the interaction between order flows and price dynamics. Furthermore, we discover both bid and ask trading volumes of the limit order book affect the price. Hence we are able to answer the first three questions proposed in the very beginning, i) How does the order flows interact with price dynamics? ii) Are the impacts on return responding to incoming ask and bid limit orders widely symmetric? iii) If not symmetric, how does the heterogeneous market impact caused by bid and ask order for various stocks affect the whole market? Our model has implied that in an LOB market, the huge sell/buy volume queued on the ask/bid side could induce strong sell/buy pressure on the market and therefore changing the price. In the following, we will focus on the last question, iv) How to measure the impact of market/limit order quantitatively?

5.5 Measuring Price Direction under Uncertainty Shock

When a large market order to buy or sell a stock arrives, the market order will automatically execute, this causes a temporary market impact. Even though sufficiently large market order immediately affects the price direction, the bid/ask sizes alone do not give enough information on price direction. To solve this, we use structural analysis proposed in section 5.3.2 to measure the persistent effect of shock in the LOB. In this section, our aim is to gain some insights into the details of the price formation and explore the existence of arbitrage opportunities.

To measure the impacts of market/limit order and whether the impacts identified by our model are temporary or robust over time, we resort to generalized impulse response analysis similar to the GI defined in (5.10). However we assume a unit shock hitting only one equation at a time, its impact on jth equation of y_t is the following,

$$\begin{array}{rcl} \delta_{jt} & : & (\delta_{1t}, \delta_{2t}, \dots, \delta_{Kt})^\top \sim e_j \\ GI(l, \delta_{jt}, \mathcal{F}_{t-1}) & = & \mathsf{E}(y_{t+l} \mid u_{jt} = \delta_{jt}, \mathcal{F}_{t-1}) - \mathsf{E}(y_{t+l} \mid \mathcal{F}_{t-1}) \end{array} \tag{5.30}$$

where $\mathsf{E}(y_{t+1} \mid u_{jt} = \delta_{jt}, \mathcal{F}_{t-1})$ represents the expectation conditional on the history $\mathcal{F}_{t-1}$ and a fixed value of j-th shock δ_{jt} on time t at horizon l. $\mathcal{F}_{t-1}$ consists of the information used to compute the conditional expectations based on bootstrap method.

Our starting point is based on market impacts regarding their own trading activities. To measure the market impacts of the order flows on price factor at a given horizon l for a stock n, the response of price factor $\Delta\tilde{p}_t^{(m)}$ are quantified by equation (5.30) when the shock δ_{jt} is treated as one of the size factors ($\Delta\tilde{s}_t^{a1(n)}$, $\Delta\tilde{s}_t^{a2(n)}$, $\Delta\tilde{s}_t^{a3(n)}$, $\Delta\tilde{s}_t^{b1(n)}$, $\Delta\tilde{s}_t^{b2(n)}$, $\Delta\tilde{s}_t^{b3(n)}$) hitting the system. With a moderate sparse structure selected by BIC after post-LASSO, we are able to identify not only the existence of significant market impact, but also the pattern of own-price market impact when w[h!]

	MSFT	T	IBM	JNJ	PFE	MRK	JPM	WFC	C
$as1$		⊖ ⊖ ⊖ ⊖ ⊖		⊕ ⊕ ⊕		⊕⊖	⊖ ⊖ ⊖	⊖ ⊖ ⊖	⊖
$as2$	⊖ ⊖ ⊖	⊕	⊕⊕		⊕			⊖⊖	⊖ ⊖ ⊖ ⊖ ⊖⊖
$as3$	⊕⊕	⊕	⊕	⊕	⊕ ⊕ ⊕		⊖	⊕⊕	⊖
$bs1$	⊕ ⊖ ⊖	⊕ ⊖ ⊖⊖		⊕	⊕	⊕⊕	⊕ ⊕ ⊕⊖	⊕⊕	⊕
$bs2$	⊖	⊖			⊕⊖	⊕	⊕		⊕
$bs3$		⊖	⊖⊖	⊖⊖	⊖	⊖⊖	⊖		
r_{size}	44%	46%	0%	14%	25%	57%	80%	78%	100%

Table 5.13. The summary of own-price market impacts.

$m = n$ and cross-price market impact when $m \neq n$. Here we use $l = 30min$ and calculate the corresponding bootstrapped GI estimation for every trading day.

We identify in total 10 days where there are significant own-price market impacts for Wells Fargo. As an example, Figure 5.6 depicts the result on 25th of July. We observe a negative correlation between the magnitude of its ask market order and price factor. It is normal for financial market in the sense that the investors will start marking down their bid price when there is a wave of sell orders coming into the order book. As expected, the price (average of bid and ask quotes) factor tends to decrease significantly after the arrival of a large ask market order. This argument holds for the case of bid market order as well. In Figure 5.7, we observe a positive market impact from bid market order on 19th of July, 2016. Both impacts can last for almost 10 minutes before the price shifts back, this gives the HF investors enough time of reaction to arbitrage opportunities.

Figure 5.8 shows the market impacts of orders posted deeper in the book for Citigroup. This implies the positive pile-on effect where larger ask order may further perpetuating a price decrease, the orders may not necessarily set at the current market price of the stock (i.e. they are not market orders, they are limit orders). The estimated market impact lasts for almost 20 minutes, the price goes up after 10 minutes because the market investors may buy trades picking up the posted volume or by cancellations on the ask side.

Table 5.13 reports the summary of significant market impacts identified by our model. For each trading day, we use ⊖ and ⊕ to represent the significant negative and positive response of price after the arrival of a market/limit order. Specifically, we define a ratio denoted as r_{size} to measure the price direction of market impacts,

$$\begin{aligned} r_{size} &= \frac{|\operatorname{sgn}(GI_t)|}{42} \\ \operatorname{sgn}(GI_t) &= \begin{cases} -1 & -GI_t(h) > Q_{0.05}(GI_t(h)) \\ 0 & |GI_t(h)| \leqslant Q_{0.05}(GI_t(h)) \\ 1 & GI_t(h) > Q_{0.05}(GI_t(h)) \end{cases} \\ t &= 1 \ldots T, \quad h = 1, \ldots, 30 \end{aligned} \tag{5.31}$$

The results suggest that the group of financial stocks is of higher r_{size} values, this may be explained by the fact that finance sector is leading the market, the history information indicates that their response of price to trading volumes is stable and thus robust for statistical arbitrage, see Hautsch and Huang (2012). The Citigroup performs well among them. Interestingly, the healthcare and technology stocks sometimes show opposite results, we notice that their prices are positively linked to ask order flows in some cases. This is because they are price-dominated stocks, i.e., they have multiple risk sources except for their own trading activity. This result is consistent with our main findings in subsection 5.4.1 and 5.4.2 where we conclude that financial stocks are size-dominated stocks and they are influencers in the system. Alternatively, the price of healthcare and technology stocks are risk receivers. Based on our methodology, it would be more profitable to invest in financial stocks for algorithm traders.

5.6 Conclusion

This paper build upon and extend current literature where the connectedness measures are often estimated by MA transformation of VAR systems and restricted to Gaussian innovations. We combine bootstrap-based generalized impulse response analysis with network construction. In this way, the network we construct relies neither on the ordering of the variables nor on the distribution of the innovations, the resulting connectedness measures is economic interpretable. Furthermore, given the HF LOB NASDAQ data, network analysis of LOB across stocks becomes interesting. Throughout the paper, we first show how network for LOB can be constructed in the presence of microstructure noise and non-synchronous trading, then we progress by focusing on the models that capture the dynamics of LOB and their influence over time. Our primary finding is that the network that involving the trading volumes is a better measure of the stock connectedness. With our methodology, we identify the significant market impact caused by the arrival of a large limit order, and order imbalance generally exists across stocks, bootstrapped market impacts can be quantified. The financial institutions are connected more closely compared with the firms come from other industry.

5.A Pre-averaging estimation

Suppose that we observe non-synchronous noisy data Y_t following,

$$Y_t = X_t + \varepsilon_t, \quad t \geqslant 0 \tag{5.32}$$

with efficient log price X_t is latent. The error term ε_t represents microstructure noise and is assumed to be independent and identically distributed with

$$\mathsf{E}(\varepsilon_t) = 0, \quad \mathsf{E}\left(\varepsilon_t^2\right) = \psi \tag{5.33}$$

The price process X_t follows a semi-martingale form, Delbaen and Schachermayer (1994),

$$X_t = X_0 + \int_0^t a_s ds + \int_0^t \sigma_s dW_s \tag{5.34}$$

where $(a_s)_{s\geqslant 0}$ is a càdlàg drift process, $(\sigma_s)_{s\geqslant 0}$ is an adapted càdlàg volatility process, $(W_s)_{s\geqslant 0}$ is a Brownian motion. In addition, we assume X_t and ε_t are independent, i.e.

$$\mathsf{E}(\varepsilon_t \mid X) = 0 \tag{5.35}$$

If one can only observe Y_i^n at discrete times t, i indexes the time points with interval length Δ_n, the returns $\Delta_i^n Y$ is thus defined as,

$$Y_i^n = Y_{i\Delta_n}, \quad \Delta_i^n Y = Y_i^n - Y_{i-1}^n, \quad i = 1, \ldots, n \tag{5.36}$$

A pre-averaging is conducted to alleviate microstructure noise and solve non-synchronicity, we follow the notations originally used by Jacod et al. (2009). The basic idea is to construct smoothing functions to diminish the impact of the noise induced by ε_t. Specifically, there is a sequence of integers denoted as k_n which satisfies,

$$\exists \theta > 0, \quad k_n \sqrt{\Delta_n} = \theta + o\left(\Delta_n^{\frac{1}{4}}\right) \tag{5.37}$$

and a continuous weight function $g : [0, 1] \mapsto \mathbb{R}$. g is piecewise C^1 with a piecewise derivative $g^{'}$, $g(0) = g(1) = 0$, and $\int_0^1 g^2(s)ds > 0$. Furthermore, the following real-valued numbers and functions are associated with function g on $\mathbb{R}_+$,

$$\begin{aligned} \psi_1 &= \int_0^1 \{g^{'}(u)\}^2 du, \quad \psi_2 = \int_0^1 \{g(u)\}^2 du \\ \Phi_1(s) &= \int_s^1 g^{'}(u) g^{'}(u-s) du, \quad \Phi_2(s) = \int_s^1 g(u) g(u-s) du \\ \Phi_{ij} &= \int_0^1 \Phi_i(s) \Phi_j(s) du, \quad i, j = 1, 2, \quad u \in [0, 1] \end{aligned} \tag{5.38}$$

Here we choose $g(x) = x \wedge (1 - x)$, as in Podolskij et al. (2009), Christensen et al. (2010) and Hautsch and Podolskij (2013). Therefore we have

$$\begin{aligned} \psi_1 &= 1, \quad \psi_2 = \frac{1}{12}, \quad \Phi_{11} = \frac{1}{6} \\ \Phi_{12} &= \frac{1}{96}, \quad \Phi_{22} = \frac{151}{80640} \end{aligned} \tag{5.39}$$

The pre-averaged returns $\overline{Y}_i^n$ associated with the weight function g are given as,

$$\begin{aligned} \overline{Y}_i^n &= \sum_{j=1}^{k_n - 1} g\left(\frac{j}{k_n}\right) \Delta_{i+j}^n Y \\ &= -\sum_{j=0}^{k_n - 1} \left\{ g\left(\frac{j+1}{k_n}\right) - g\left(\frac{j}{k_n}\right) \right\} Y_{i+j}^n, \quad i = 0, \ldots, n - k_n + 1 \end{aligned} \tag{5.40}$$

The window size k_n defined in equation (5.37) is chosen of $\mathcal{O}\left(\sqrt{\frac{1}{\Delta_n}}\right)$, balance the noise $\overline{\varepsilon}_i^n = \mathcal{O}_p\left(\sqrt{\frac{1}{k_n}}\right)$ and the efficient price $\overline{X}_i^n = \mathcal{O}_p\left(\sqrt{k_n \Delta_n}\right)$.

5.B Bootstrap-based multistep forecast methods

Here we describe the computational steps to obtain the $\mathsf{E}(y_{t+1}|u_{jt} = \delta_{jt}, \mathcal{F}_{t-1})$, GI, GFEVD via Bootstrap method, more details can be found in Koop et al. (1996), Lanne and Nyberg (2016), Teräsvirta et al. (2010).

1. Denote $\mathcal{F}_{t-1}$ as all the information prior to Y_t, and select a forecast horizon h.
2. Randomly sample N_B vectors of shocks $(\delta_{1t}, \delta_{2t}, \dots, \delta_{Kt})^\top$ from the residuals of estimated model,

$$\delta_{jt} : (\delta_{1t}, \delta_{2t}, \dots, \delta_{Kt})^\top \sim \hat{u}^\star_{jt} e_j \tag{5.41}$$

$$\hat{u}^\star_{jt} = Y_t - \left(\hat{A}_1, \hat{A}_2, \dots, \hat{A}_p\right) \left(Y^\top_{t-1}, Y^\top_{t-2}, \dots, Y^\top_{t-p}\right)^\top = Y_t - g(Y_{t-1}) \tag{5.42}$$

3. Compute conditional multistep forecast $\mathsf{E}(y_{t+l}|\mathcal{F}_{t-1})$,

$$\begin{aligned} f_{t,0} &= g(Y_{t-1}) \\ f_{t,1} &= \mathsf{E}[Y_{t+1} \mid \mathcal{F}_{t-1}] = \mathsf{E}[g(f_{t,0} + \hat{u}^\star_t) \mid \mathcal{F}_{t-1}] \\ f_{t,2} &= \mathsf{E}[Y_{t+2} \mid \mathcal{F}_{t-1}] = \mathsf{E}[g(f_{t,1} + \hat{u}^\star_{t+1}) \mid \mathcal{F}_{t-1}] \\ &\dots \end{aligned} \tag{5.43}$$

   with $\hat{u}^\star_{t+l}, l = 1, \dots, h$ are independent draws with replacement from the set of residuals $\{\hat{u}_{t+l}\}_{t=1}^T$ over the sample period.

4. Repeat steps 3 for all N_B vectors of estimated innovations with bootstrap methods, iterating on the estimated model,

$$\begin{aligned} fb_{t,1} &= \frac{1}{N_B} \sum_{i=1}^{N_B} g(f_{t,0} + \hat{u}^{\star(i)}_t) \\ fb_{t,2} &= \frac{1}{N_B} \sum_{i=1}^{N_B} g(g(f_{t,0} + \hat{u}^{\star(i)}_t) + \hat{u}^{\star(i)}_{t+1}) \\ &\dots \end{aligned} \tag{5.44}$$

5. By the same logic, we compute $\mathsf{E}(y_{t+l} \mid u_{jt} = \delta_{jt}, \mathcal{F}_{t-1})$ when the shock is given as $\delta_{jt} = \hat{u}^\star_{jt} e_j$,

$$\begin{aligned} f_{t,0} &= g(Y_{t-1}) \\ f_{t,1} &= \mathsf{E}[Y_{t+1} \mid \mathcal{F}_{t-1}] = \mathsf{E}[g(f_{t,0} + \hat{u}^\star_{jt} e_j) \mid \mathcal{F}_{t-1}] \\ f_{t,2} &= \mathsf{E}[Y_{t+2} \mid \mathcal{F}_{t-1}] = \mathsf{E}[g(f_{t,1} + \hat{u}^\star_{j,t+1} e_j) \mid \mathcal{F}_{t-1}] \\ &\dots \end{aligned} \tag{5.45}$$

with $\hat{u}^{\star}_{j,t+l}, l = 1, \ldots, h$ are independent draws with replacement from the set of residuals $\{\hat{u}_{j,t+l}\}_{t=1}^{T}$ over the sample period.

6. Repeat steps 5 for all N_B vectors of estimated innovations with bootstrap methods, iterating on the estimated model,

$$\begin{aligned} fb_{t,1} &= \frac{1}{N_B}\sum_{i=1}^{N_B} g(f_{t,0} + \hat{u}^{\star(i)}_{jt} e_j) \qquad (5.46) \\ fb_{t,2} &= \frac{1}{N_B}\sum_{i=1}^{N_B} g(g(f_{t,0} + \hat{u}^{\star(i)}_{jt} e_j) + \hat{u}^{\star(i)}_{j,t+1} e_j) \\ &\ldots \end{aligned}$$

7. Plug in the GI function

$$GI(l, \delta_{jt}, \mathcal{F}_{t-1}) = \mathsf{E}(y_{t+l} \mid u_{jt} = \delta_{jt}, \mathcal{F}_{t-1}) - \mathsf{E}(y_{t+l} \mid \mathcal{F}_{t-1}) \qquad (5.47)$$

to obtain the relative contribution of a shock δ_{jt} to the i-th variable with horitzon h at time t,

$$\lambda_{ij,\mathcal{F}_{t-1}}(h) = \frac{\sum_{l=0}^{h} GI(l, \delta_{jt}, \mathcal{F}_{t-1})_i^2}{\sum_{j=1}^{K}\sum_{l=0}^{h} GI(l, \delta_{jt}, \mathcal{F}_{t-1})_i^2}, \quad i, j = 1, \ldots, K \qquad (5.48)$$

8. Repeat steps 2-6 for all histories.

9. Construct table 5.3 using averaged $\lambda_{ij,\mathcal{F}_{t-1}}(h)$ generated from step 7.

If there is a unit shock,

$$\delta_{jt} \ : \ (\delta_{1t}, \delta_{2t}, \ldots, \delta_{Kt})^{\top} \sim e_j \qquad (5.49)$$

then we can simiply replace $\hat{u}^{\star}_{jt} e_j$ of (5.41) with e_j of (5.49), and repeat the steps from 1 to 6 stated above, the generalized impulse response can be calculated based on (5.47), i.e.

$$GI(l, \delta_{jt}, \mathcal{F}_{t-1}) = \mathsf{E}(y_{t+l} \mid u_{jt} = \delta_{jt}, \mathcal{F}_{t-1}) - \mathsf{E}(y_{t+l} \mid \mathcal{F}_{t-1}) \qquad (5.50)$$

We should note that if K is extremely large in empirical study, the denominator of equation (5.48) might be unnecessarily large due to accumulated noise caused by the large amount of irrelevant variables. Therefore one more step of prescreening is preferred to filter out less relevant variables.

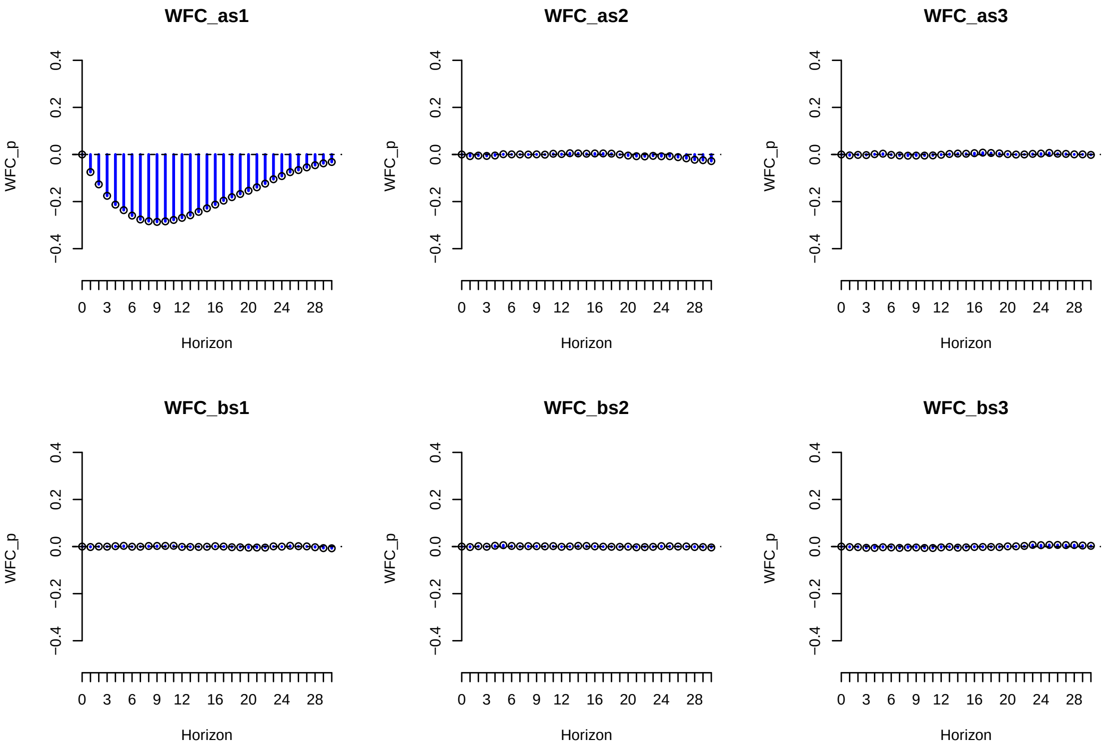


Figure 5.6. Own-price market impact of WFC (Wells Fargo) on 25th of July, 2016.

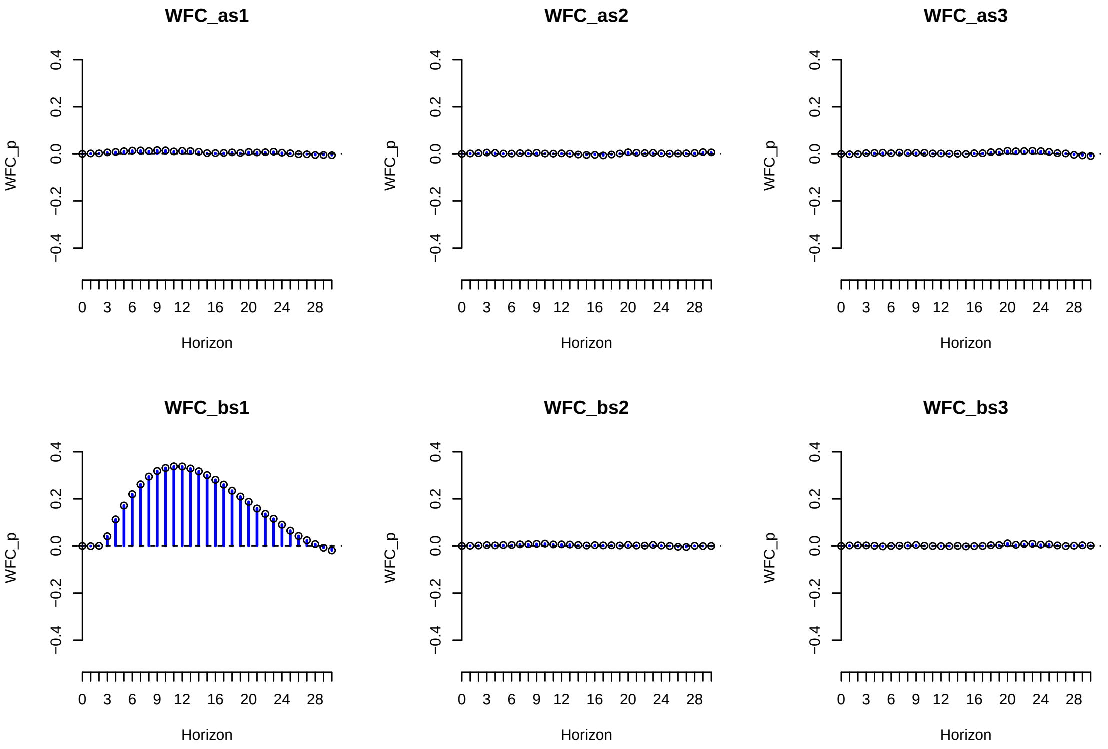


Figure 5.7. Own-price market impact of WFC (Wells Fargo) on 19th of July, 2016.

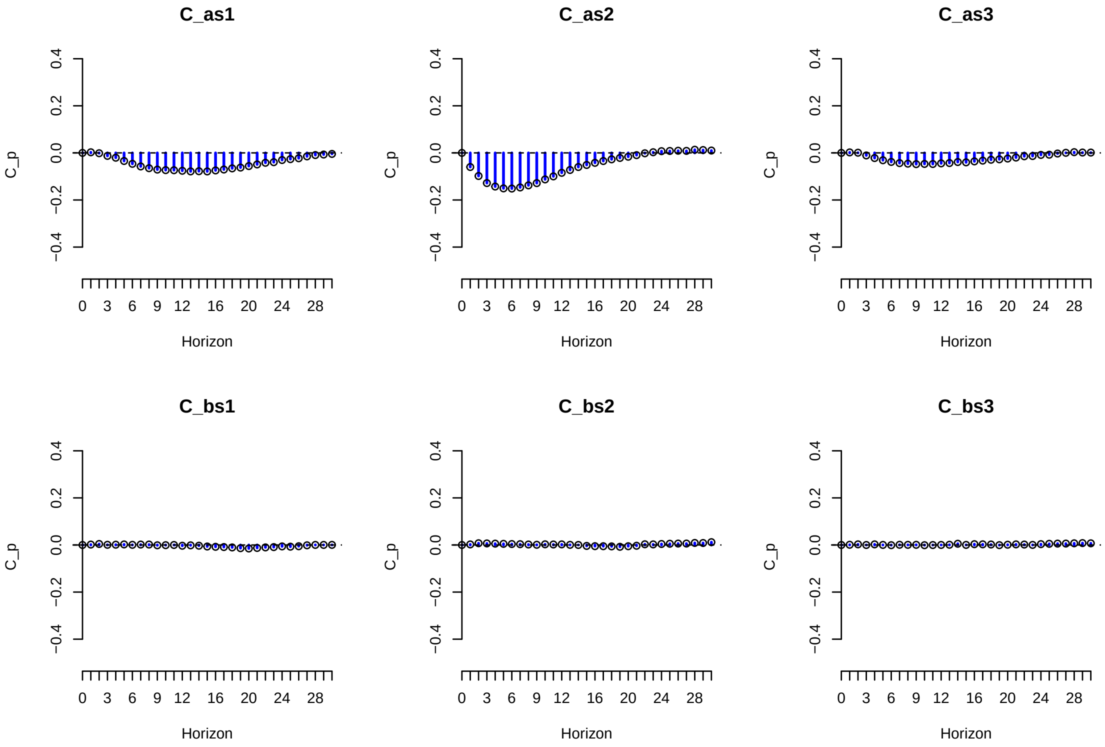


Figure 5.8. The bootstrapped market impact of Citigroup on 1st of June, 2016.

Bibliography

Ahn, S. K. and Reinsel, G. C. (1990). Estimation for partially nonstationary multivariate autoregressive models. *Journal of the American Statistical Association*, 85(411):813–823.

Aït-Sahalia, Y., Mykland, P. A., and Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial studies*, 18(2):351–416.

Andersen, T. G. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International economic review*, pages 885–905.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2000). Exchange rate returns standardized by realized volatility are (nearly) gaussian. Technical report, National Bureau of Economic Research.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American statistical association*, 96(453):42–55.

Anderson, T. (2002). Reduced rank regression in cointegrated models. *Journal of Econometrics*, 106(2):203–216.

Andrew, G. and Gao, J. (2007). Scalable training of l 1-regularized log-linear models. In *Proceedings of the 24th international conference on Machine learning*, pages 33–40. ACM.

Bandi, F. M. and Russell, J. R. (2006). Separating microstructure noise from volatility. *Journal of Financial Economics*, 79(3):655–692.

Bandi, F. M. and Russell, J. R. (2008). Microstructure noise, realized variance, and optimal sampling. *The Review of Economic Studies*, 75(2):339–369.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76(6):1481–1536.

Barndorff-Nielsen, O. E. and Shephard, N. (2001). Non-gaussian ornstein–uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):167–241.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.*, 43(4):1535–1567.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, pages 2577–2604.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732.

Bloomfield, R., O?hara, M., and Saar, G. (2005). The ?make or take? decision in an electronic market: Evidence on the evolution of liquidity. *Journal of Financial Economics*, 75(1):165–199.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351.

Cavaliere, G., Angelis, L. D., Rahbek, A., and Robert Taylor, A. M. (2014). A comparison of sequential and information-based methods for determining the co-integration rank in heteroskedastic var models. *Oxford Bulletin of Economics and Statistics*, page forthcoming.

Cavaliere, G., Rahbek, A., and Taylor, A. M. R. (2012). Bootstrap determination of the co-integration rank in vector autoregressive models. *Econometrica*, 80(4):1721–1740.

Chao, J. C. and Phillips, P. C. (1999). Model selection in partially nonstationary vector autoregressive processes with reduced rank structure. *Journal of Econometrics*, 91(2):227 – 271.

Chen, X., Xu, M., and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *Ann. Statist.*, 41(6):2994–3021.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Comparison and anti-concentration bounds for maxima of gaussian random vectors.

Christensen, K., Kinnebrock, S., and Podolskij, M. (2010). Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. *Journal of Econometrics*, 159(1):116–133.

Chung, K. and Williams, R. (1990). *Introduction to Stochastic Integration*. Probability and Its Applications. Springer New York.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics. Springer New York.

Delbaen, F. and Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing. *Mathematische annalen*, 300(1):463–520.

Demirer, M., Diebold, F. X., Liu, L., and Yılmaz, K. (2017). Estimating global bank network connectedness. Technical report, National Bureau of Economic Research.

Diebold, F. X. and Yılmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics*, 182(1):119–134.

Eisler, Z., Bouchaud, J.-P., and Kockelkoren, J. (2012). The price impact of order book events: market orders, limit orders and cancellations. *Quantitative Finance*, 12(9):1395–1419.

Engle, R. and Granger, C. (1987). Co-integration and error correction: representation, estimation and testing. *Econometrica*, 55:257–276.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Foucault, T., Kadan, O., and Kandel, E. (2005). Limit order book as a market for liquidity. *The review of financial studies*, 18(4):1171–1217.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *ArXiv e-print*, 1001.0736.

Gao, C., Ma, Z., Ren, Z., and Zhou, H. H. (2015). Minimax estimation in sparse canonical correlation analysis. *Ann. Statist.*, 43(5):2168–2197.

Golub, G. and Van Loan, C. (2013). *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press.

Hamilton, J. (1994). *Time series analysis.* Princeton Univ. Press, Princeton, NJ.

Handa, P., Schwartz, R., and Tiwari, A. (2003). Quote setting and price formation in an order driven market. *Journal of financial markets*, 6(4):461–489.

Hautsch, N. and Huang, R. (2012). The market impact of a limit order. *Journal of Economic Dynamics and Control*, 36(4):501–522.

Hautsch, N. and Podolskij, M. (2013). Preaveraging-based estimation of quadratic variation in the presence of noise and jumps: theory, implementation, and empirical evidence. *Journal of Business & Economic Statistics*, 31(2):165–183.

Jacod, J., Li, Y., Mykland, P. A., Podolskij, M., and Vetter, M. (2009). Microstructure noise in the continuous case: the pre-averaging approach. *Stochastic processes and their applications*, 119(7):2249–2276.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2-3):231 – 254.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica*, 59(6):pp. 1551–1580.

Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models.* Oxford University Press.

Johnson, W. and Lindenstrauss, J. (2001). *Handbook of the Geometry of Banach Spaces.* Number Bd. 1 in Handbook of the Geometry of Banach Spaces. Elsevier Science.

Kavajecz, K. A. and Odders-White, E. R. (2004). Technical analysis and liquidity provision. *Review of Financial Studies*, 17(4):1043–1071.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):pp. 1356–1378.

Kock, A. B. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2):325–344.

Koh, K., Kim, S.-J., Boyd, S., and Lin, Y. (2007). An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine Learning Research*, 2007.

Koop, G., Pesaran, M. H., and Potter, S. M. (1996). Impulse response analysis in nonlinear multivariate models. *Journal of econometrics*, 74(1):119–147.

Kosorok, M. R. and Ma, S. (2007). Marginal asymptotics for the large p, small n paradigm: With applications to microarray data. *Ann. Statist.*, 35(4):1456–1486.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.*, 40(2):694–726.

Lanne, M. and Nyberg, H. (2016). Generalized forecast error variance decomposition for linear and nonlinear multivariate models. *Oxford Bulletin of Economics and Statistics*, 78(4):595–603.

Li, H., Li, Q., and Shi, Y. (2017). Determining the number of factors when the number of factors can increase with sample size. *Journal of Econometrics*, 197(1):76–86.

Li, R.-C. (1994). On perturbations of matrix pencils with real spectra. *Mathematics of Computation*, 62(205):231–265.

Liao, Z. and Phillips, P. C. (2015). Automated estimation of vector error correction models. *Econometric Theory*, 31(03):581–646.

Liski, E., Nordhausen, K., Oja, H., and Ruiz-Gazen, A. (2016). Combining linear dimension reduction subspaces. *proceedings of ICORS 2015*.

Liu, W., Xiao, H., and Wu, W. B. (2013). Probability and moment inequalities under dependence. *Statistica Sinica*, 23(3):1257–1272.

Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Springer Publishing Company, Incorporated.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097.

Onatski, A. and Wang, C. (2018). Alternative asymptotics for cointegration tests in large vars. Technical report, Cambridge-INET Working Paper Series No: 2016/07.

Parlour, C. A. and Seppi, D. J. (2003). Liquidity-based competition for order flow. *The Review of Financial Studies*, 16(2):301–343.

Pesaran, H. H. and Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. *Economics letters*, 58(1):17–29.

Phillips, P. C. (2014). Optimal estimation of cointegrated systems with irrelevant instruments. *Journal of Econometrics*, 178(Part 2):210 – 224. Recent Advances in Time Series Econometrics.

Podolskij, M., Vetter, M., et al. (2009). Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps. *Bernoulli*, 15(3):634–658.

Roşu, I. (2009). A dynamic model of the limit order book. *The Review of Financial Studies*, 22(11):4601–4641.

Signoretto, M. and Suykens, J. (2012). Convex estimation of cointegrated VAR models by a nuclear norm penalty. *IFAC Proceedings*, 45(16):95 – 100.

Stewart, G. W. (1984). Rank degeneracy. *SIAM Journal on Scientific and Statistical Computing*, 5(2):403–413.

Stewart, G. W. and Sun, J. (1990). *Matrix Perturbation Theory.* Academic Press.

Teräsvirta, T., Tjostheim, D., Granger, C. W., et al. (2010). Modelling nonlinear economic time series. *OUP Catalogue.*

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):pp. 267–288.

Toda, H. Y. and Phillips, P. C. (1993). Vector autoregressions and causality. *Econometrica: Journal of the Econometric Society*, pages 1367–1393.

Wang, H. and Leng, C. (2008). A note on adaptive group lasso. *Computational Statistics & Data Analysis*, 52(12):5277–5286.

Wang, Z. and Bessler, D. A. (2005). A monte carlo study on the selection of cointegrating rank using information criteria. *Econometric Theory*, 21:593–620.

Wei, F. and Huang, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli*, 16(4):1369–1384.

Wilms, I. and Croux, C. (2016). Forecasting Using Sparse cointegration. *International Journal of Forecasting*, 32:1256–1267.

Wu, W. B. (2007). Strong invariance principles for dependent random variables. *Ann. Probab.*, 35(6):2294–2320.

Wu, W.-B. and Wu, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electron. J. Statist.*, 10(1):352–379.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67.

Zhang, L. (2011). Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics*, 160(1):33–47.

Zhang, L. et al. (2006). Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli*, 12(6):1019–1043.

Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100(472):1394–1411.

Zhang, R., Robinson, P., and Yao, Q. (2018). Identifying cointegration by eigenanalysis. *Journal of the American Statistical Association*, 0(0):1–12.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):pp. 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

List of Figures

List of Tables