Large Spillover Networks of Nonstationary Systems

Shi Chen* Melanie Schienle†

Abstract

This paper proposes a vector error correction framework for constructing large consistent spillover networks of nonstationary systems grounded in the network theory of Diebold and Yılmaz (2014). We aim to provide a tailored methodology for the large non-stationary (macro)economic and financial system application settings avoiding technical and often hard to verify assumptions for general statistical high-dimensional approaches where the dimension can also increase with sample size. To achieve this, we propose an elementwise Lasso-type technique for consistent and numerically efficient model selection of VECM, and relate the resulting forecast error variance decomposition to the network topology representation. We also derive the corresponding asymptotic results for model selection and network estimation under standard assumptions. Moreover, we develop a refinement strategy for efficient estimation and show implications and modifications for general dependent innovations. In a comprehensive simulation study, we show convincing finite sample performance of our technique in all cases of moderate and low dimensions. In an application to a system of FX rates, the proposed method leads to novel insights on the connectedness and spillover effects in the FX market among the OECD countries.

JEL classification: C3, C5, F3

Keywords: network, connectedness, cointegration, VECM, exchange rates, adaptive Lasso, nonstationary, spillover

*Karlsruhe Institute of Technology, Chair of Econometrics and Statistics, Blücherstr.17, 76185 Karlsruhe, Germany. Email: shi.chen@kit.edu.

†Karlsruhe Institute of Technology, Chair of Econometrics and Statistics, Blücherstr.17, 76185 Karlsruhe, Germany. Email: melanie.schienle@kit.edu.

1 Introduction

In recent years, the analysis of networks over time has become central for understanding and estimating complex systems in macroeconomics and finance. Generally, links between the components of the system might act as the carriers of systemic risk transmission and thus identifying the connectedness structure has become research focus in order to uncover the spillover effects. For example, Billio et al. (2012) use a Granger causal network to measure systemic risk across and within different parts of the financial sector; in the framework of vector autoregressive (VAR) model, Diebold and Yilmaz (2009, 2012) and later Diebold and Yılmaz (2014) propose a volatility spillover network using the generalized variance decomposition of Pesaran and Shin (1998).

However many economic and financial systems are dynamic, multi-dimensional and often contain a large number of non-stationary potentially cointegrated components, the standard VAR setting does not consider a potential cointegration structure. To handle such multivariate time-series, we use the VECM as introduced in Engle and Granger (1987). While already for settings greater than dimension two, standard econometric techniques (Johansen, 1988, 1991; Xiao and Phillips, 1999; Hubrich et al., 2001; Boswijk et al., 2015) often fail to provide accurate, testable and computationally tractable estimates, there has emerged a recent literature on high-dimensional estimation (Liang and Schienle, 2019; Zhang et al., 2018) in this context. The generality of the latter approaches, however, comes with a set of technical assumptions which are hard to verify in practice and lacking asymptotic distributional results which are key for inference. Thus in particular in view of many macroeconomic applications, there is a need for easy to use practically feasible techniques with available asymptotic distributions for cases where cross-sectional dimensions are moderately large, i.e. large but not expanding with sample size. We show that for such settings, not only assumptions simplify and asymptotic confidence regions exist, but also novel tailored procedures can be designed. Such techniques would not be feasible in the fully high-dimensional setup, but allow for a more refined identification of non-zero elements in the moderate dimensional model.

In our setting, the above VECM estimation results can be associated with several network structures such as Dahlhaus (2000), Eichler (2007), Eichler (2012) and Diebold and Yılmaz (2014). Here we focus on one particular structure, the DY network, following the work of Diebold and Yılmaz (2014). To estimate the VECM, we propose an adaptive shrinkage method that simultaneously allows for model choice and direct estimation. 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 actually occur in the system. In the same way in practice, a small number of VAR lags are considered sufficient for a parsimonious model specification, i.e. within a maximum lag range only a small number of effective lags are relevant, which are not required to be consecutive. In this sense, the problem is assumed to be "sparse". In contrast to a fully high-dimensional set-up, this "sparsity" is not necessary for consistent model identification but only increases the numerical efficiency and thus the feasibility of our procedure. We show consistency of the variable selection by the proposed Lasso-VECM estimator and derive its asymptotic properties for inference. For more efficient estimation in particular in cases with a small sample for a large cross-section dimension, we provide a refined estimation strategy and derive its statistical properties. Our presented methods here are tailored to the moderate fixed-dimensional case where elementwise adaptive lasso penalization is still numerically feasible. For such cases which are prevalent in macroeconomic applications, the techniques can identify not only the cointegration rank and lag consistently but also non-zero elements in the structure of the cointegration space. A simulation study shows the effectiveness of the proposed techniques in finite samples. In addition, we conduct an empirical study for quarterly floating exchange (FX) data for a system of 17 OECD countries. There is a sizable literature suggesting that, especially at short horizons, a random walk forecast of the exchange rate generally outperforms alternative models (such as Meese and Rogoff (1983)). This indicates that the FX series contain nonstationary dynamics and VECM is required to handle such large system. Our FX application illustrates that such refinements can make a difference in practice.

Our theoretical work builds on the vast literature of VECM as summarized e.g. in Lütkepohl (2007) as well as on results for adaptive Lasso techniques as in Zou (2006) and Yuan and Lin (2006) and Medeiros and Mendes (2016). More recently, our technique also relates to the work of Kock and Callot (2015), Barigozzi and Brownlees (2019) which use Lasso for model determination in a stationary high-dimensional VAR context but cannot handle nonstationary components. For non-stationary time series, there also exists some empirical and simulation work employing penalizing algorithms for VECM without proofs, see e.g. Signoretto and Suykens (2012), Wilms and Croux (2016). Some theoretical results for a nonlinear penalization criterion in fixed-dimensional VECM have been derived by Liao and Phillips (2015). However their theoretical results hold only for real eigenvalues but complex eigenvalues will occur in the numerical and empirical examples. Our proposed linear Lasso approach, however, does not require a symmetric cointegration matrix and thus provides a feasible solution for general moderate to high dimensional settings. These non-symmetric cases are the rule rather than the exception where eigenvalue based methods have not only feasibility problems but fail to get any real-valued solution at all. In contrast to the general but only group-wise rough high-dimensional shrinkage in Liang and Schienle (2019), the presented moderate dimensional technique can identify non-zero

elements in the cointegration space and avoid technical and hard to verify eigenvalue type assumptions. For applications, this can be key to augmented forecasting results as illustrated in the studied FX case. The paper is also related to the high-dimensional factor model without explicit VECM structure in Lam and Yao (2012), Zhang et al. (2018) and the high-dimensional distributional results by random matrix theory in Onatski and Wang (2018). Compared to Lam and Yao (2012), Zhang et al. (2018), we incorporate standard factor model with VECM structure and apply Lasso to determine the rank. In contrast to Onatski and Wang (2018), our focus in on consistent model selection rather than the distribution of eigenvalues.

The paper is organized as follows. Section 2 presents the model setup. Section 3 provides the determination of underlying dynamics. We first show the lasso-type technique for consistent and numerically efficient model selection of VECM. Second, we give the main asymptotic results on model selection consistency and derives the asymptotic distribution for estimates. We also show strategies for refined estimation and derive results when the error terms are weakly dependent. Section 5 presents comprehensive simulation results, as well as the empirical findings for FX rates. All proofs are contained in the Appendix.

2 Model setup

We consider a VECM setup with $\{Y_t\}$ is a nonstationary m-dimensional $I(1)$ process, $\Delta Y_t = Y_t - Y_{t-1}$ is stationary. Suppose general structure of the true process $\{Y_t\}$ follows

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

for $t = 1, \ldots, T$. Π is an $m \times m$ matrix of rank r with $0 \leq r < m$, marking the number of cointegration relations in the system. Π can be further decomposed as $\Pi = \alpha\beta'$, where β marks the r long-run cointegrating relations and α is a loading matrix of rank r. Without loss of generality, we set β as orthogonal, i.e. $\beta'\beta = I_r$. Then the decomposition $\Pi = \alpha\beta'$ is unique up to an orthonormal H, the cointegration relations β are identified up to rotation. 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, $B_{p+1}, \ldots, B_P$ are all zero matrices. Additionally we assume that $m/r = c_1$ and $P/p = c_2$ with $c_1, c_2 \gg 1$, i.e. c_1, c_2 are substantially exceeding 1, meaning that the number of cointegration relations is small relative to m as well as the effective lag length p is much smaller than P. Note that in contrast to a fully high-dimensional set-up, this "sparsity" type assumption is not necessary for consistent model identification but only increases the numerical efficiency and thus the feasibility of our procedure.

For the error term u_t, we first employ a standard white noise assumption to focus on the key aspects of our Lasso selection procedure while keeping technical results simple. Later in Section 3.3.2, we show how this *i.i.d* assumption can be relaxed allowing linear forms of weak dependence. Though, we show that such a general setting requires changes in the Lasso procedure and leads to different statistical properties of the modified technique losing the elementwise advantages. Generally, the normality in the following Assumption 2.1 is not crucial and can be further relaxed to only moment assumptions at the price of more involved technical arguments which, however, are not specific to our VECM set-up.

Assumption 2.1. *The error term u_t is i.i.d. $\mathcal{N}(0, \Sigma_u)$ where Σ_u is a symmetric, positive definite $m \times m$ matrix.*

Following the DY-network, we rely on the variation decomposition tool to evaluate the effect of a shock in one system variable. The network literature generally characterizes systemic risk spillover effects as connectedness obtained from a generalized forecast error variance decomposition (FEVD) of an underlying VAR system. As cointegrated variables can be generated by a VAR process, rearranging terms in (1) then gives the following VAR$(P+1)$ process in levels

$$\begin{aligned} Y_t &= (I_m + B_1 + \alpha\beta')Y_{t-1} + (B_2 - B_1)Y_{t-2} + \dots \\ &\quad + (B_P - B_{P-1})Y_{t-P} - B_P Y_{t-P-1} + u_t \end{aligned} \tag{2}$$

and we write model (2) in companion form as

$$V_t = AV_{t-1} + \omega_t \tag{3}$$

where $V_t = [Y_t', Y_{t-1}' \dots, Y_{t-p}']'$ and $\omega_t = [u_t', 0, \dots, 0]'$ and

$$A = \begin{bmatrix} I_m + B_1 + \alpha\beta' & B_2 - B_1 & \cdots & B_p - B_{p-1} & -B_p \\ I_m & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_m & 0 \end{bmatrix}. \tag{4}$$

We also use the corresponding infinite moving average (MA) representation of the system (3) in the form

$$V_t = \sum_{j=0}^{\infty} A^j \omega_{t-j} \quad \text{or} \quad Y_t = \sum_{j=0}^{\infty} \Phi_j u_{t-j} \tag{5}$$

where the jth MA coefficient matrices Φ_j are the elements of the upper left-hand $(m \times m)$ block of A^j, with $\Phi_0 = I_m$. Using the framework proposed by Koop et al. (1996) and Pesaran and Shin (1998), we specify the scaled generalized impulse response function as $IRF(j,h) = \sigma_{jj}^{-\frac{1}{2}} \Phi_h \Sigma_u e_j$ which measures the effect of one standard error shock to the jth equation at time t on expected values of Y at time $t+h$. σ_{jj} is the standard deviation of the innovation term in j-th equation. e_i is a selection vector with unity as its i-th element and zeros elsewhere. Finally, the H-step ahead generalized forecast error variance decomposition (GFEVD) $\theta_{ij}(H)$ of elements i and j is given by

$$\theta_{ij}(H) = \frac{\sigma_{jj}^{-1} \sum_{h=0}^{H-1} \left(e_i' \Phi_h \Sigma_u e_j\right)^2}{\sum_{h=0}^{H-1} \left(e_i' \Phi_h \Sigma_u \Phi_h' e_i\right)} \tag{6}$$

The DY-network works with the following $\tilde{\theta}_{ij}(H)$ which are normalized by row sum for easier interpretability

$$\tilde{\theta}_{ij}(H) = \frac{\theta_{ij}(H)}{\sum_{j=1}^{m} \theta_{ij}(H)} \tag{7}$$

Note that by construction (7) we have $\sum_{j=1}^{m} \tilde{\theta}_{ij}(H) = 1$. Accordingly, for each node i in the network we work with the following quantities as in the DY-network literature. We denote the pairwise directional connectedness $C_{i \leftarrow j}^H$ from j to i by

$$C_{i \leftarrow j}^H = \tilde{\theta}_{ij}(H) \tag{8}$$

Moreover, aggregating all effects of component i on other elements in the system, we call the total directional connectedness "to" $C_{i \leftarrow \bullet}^H$ (others to i) given by $C_{i \leftarrow \bullet}^H = C_{to,i}^H = \sum_{j=1}^{m} \tilde{\theta}_{ij}(H)/m, \text{ for } i \neq j$. Analogously, the total directional connectedness "from" $C_{\bullet \leftarrow i}^H$ (from i to others) is defined as $C_{\bullet \leftarrow i}^H = C_{from,i}^H = \sum_{i=1}^{m} \tilde{\theta}_{ij}(H)/m, \text{ for } j \neq i$. Each column refers to a variable that transmits the shock while the rows refer to respective variables that receive the transmitted shock.

Moreover the net total directional connectedness $C_{net,i}^H$ measures the direction and magnitude of the net spillover impacts of node i in the system as

$$C_{net,i}^H = C_{from,i}^H - C_{to,i}^H = C_{\bullet \leftarrow i}^H - C_{i \leftarrow \bullet}^H \tag{9}$$

and the total connectedness is given as

$$C_{total}^H = \sum_{i,j=1}^{m} \tilde{\theta}_{ij}(H)/m, \text{ for } i \neq j \tag{10}$$

Estimates of all connectedness measures are obtained by using the respective plug-in estimates in the GFEVD (6).

3 Determination of underlying dynamics

3.1 Model determination

For the detection of network spillovers in large nonstationary systems from forecast error variance decompositions, the determination and pre-estimation of the underlying dynamics is key. Note while direct VAR estimation of cointegrated systems in (2) is consistent, it is well-known that pre-detection of the cointegration parameters in the VECM specification (1) yields finite sample advantages in forecasting (see e.g. Engle and Yoo (1987)). For this, we propose a sparse and thus numerically efficient Lasso-type VECM determination technique which scales to general larger systems.

This section contains two parts: we first derive the Lasso objective function for cointegrating rank selection and estimation. Then we show the determination of the lag order in a similar manner. Therefore the model specification 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 $||\cdot||_F$ and $\to_d$ denotes convergence in distribution.

In addition to the error term assumption 2.1, our analysis 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 3.1. *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.*

For estimation purposes, we rewrite the general VECM defined in (1) in matrix notation

$$\Delta Y = \Pi Y_{-1} + B\Delta X + U \tag{11}$$

where $\Delta Y = [\Delta Y_1, \ldots, \Delta Y_T]$, $Y_{-1} = [Y_0, \ldots, Y_{T-1}]$, $B = [B_1, \ldots, B_P]$, $\Delta X = [\Delta X_0, \ldots, \Delta X_{T-1}]$

with $\Delta X_{t-1} = \left[\Delta Y'_{t-1}, \ldots, \Delta Y'_{t-P}\right]'$ and $U = [u_1, \ldots, u_T]$. W.l.o.g, $Y_k = 0$ for $k \leq 0$. Moreover, we denote $\Gamma_t = [Y'_{t-1}\beta, \Delta Y'_{t-1}, \ldots, \Delta Y'_{t-P}]'$. Under Assumptions 2.1 and 3.1, it holds by Lemma 1 in Toda and Phillips (1993)

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{[Ts]} \Gamma_t \rightarrow_p B_\Gamma(s) \tag{12}$$

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

$$\Sigma_{\Gamma\Gamma} = \begin{pmatrix} \Sigma_{z1z1} & \Sigma_{z1\Delta x} \\ \Sigma_{\Delta xz1} & \Sigma_{\Delta x\Delta x} \end{pmatrix} \tag{13}$$

Denote the LS estimate for (11) as $[\widetilde{\Pi}_{ls}, \widetilde{B}_{ls}]$, thus we obtain the consistent estimate of Σ_u as $\widetilde{\Sigma}_u = \frac{1}{T-mP+1}(\Delta Y - \widetilde{\Pi}_{ls}Y_{-1} - \widetilde{B}_{ls}\Delta X)(\Delta Y - \widetilde{\Pi}_{ls}Y_{-1} - \widetilde{B}_{ls}\Delta X)'$ (see e.g. Lütkepohl, 2007).

For model selection, we disentangle the joint lag-rank selection problem by employing the Frisch-Waugh-idea in (11). Thus we obtain two independent criteria for lag and rank choice which can be computed separately.

Rank determination For rank selection, the partial LS 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 $\Delta Y M$ on $Y_{-1}M$ with $M = I_T - \Delta X'(\Delta X \Delta X')^{-1}\Delta X$. Thus it is

$$\widetilde{\Pi} = \left(\Delta Y M Y'_{-1}\right)\left(Y_{-1} M Y'_{-1}\right)^{-1} \tag{14}$$

and we show that $\widetilde{\Pi}$ is a consistent estimate for Π in Lemma A.1 in the Appendix.

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 (11) 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

[1]For $m \geq r$, 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 $M'M_\perp = 0$.

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 criterion. Likewise, the determination of the cointegration rank in (11) 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 moderate-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 do so, we require a pre-estimate for β, which we obtain from the following QR-decomposition

$$\begin{aligned} \widetilde{\Pi} &= \widetilde{R}' \widetilde{S}' \\ &= \begin{bmatrix} \widetilde{R}'_{1,m\times r} & \widetilde{R}'_{2,m\times(m-r)} \end{bmatrix} \begin{bmatrix} \widetilde{S}'_{1,r\times m} \\ \widetilde{S}'_{2,(m-r)\times m} \end{bmatrix} \end{aligned} \tag{15}$$

where $\widetilde{S}$ is an orthonormal matrix, i.e. $\widetilde{S}'\widetilde{S} = I$. $\widetilde{R}$ is an upper triangular matrix [2]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.[3] 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 the following Lemma where we derive convergence results of the QR-decomposition components $\widetilde{R}$ and $\widetilde{S}$ from the least

[2]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.

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

squares pre-estimate $\widetilde{\Pi}'$.

Lemma 3.1. *Let Assumptions 2.1 and 3.1 hold for* $\widetilde{\Pi}$ *in* (14). *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* (15) *of* $\widetilde{\Pi}'$ *defined in* (14). *Let* β *be orthonormal and* H *be a* $(r \times r)$*-orthonormal matrix.*

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

where $\frac{1}{T}\beta' Y_{-1} M Y_{-1}' \beta \to_p \Sigma_{z1z1.\Delta x}$ *and* $\Sigma_{z1z1.\Delta x}$ *is defined as in Lemma A.1.*

Thus from Lemma A.1 and 3.1, we can construct a corresponding adaptive Lasso procedure. Hence components $\hat{R}(i,j)$ of $\hat{R}$ minimize the following criterion over all $R(i,j)$ for $i,j = 1,\ldots,m$

$$\|vec(\Delta Y M) - (M Y_{-1}' \widetilde{S} \otimes I_m) vec(R')\|^2_{I_T \otimes \Sigma_u^{-1}} + \sum_{i,j=1}^{m} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^\gamma} |R(i,j)| \tag{16}$$

where $\widetilde{R}(i,j)$ is from the QR-decomposition of $\widetilde{\Pi}'$ in the partial model (14). 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 $\hat{R}'$.

Lag order determination Likewise, for independent lag selection, the effect of the nonstationary term Y_{-1} in (11) must be filtered out in ΔY and ΔX for unbiased estimation in the partial model via regression of $\Delta Y C$ on $\Delta X C$ 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 $k = 1,\ldots,P$ and $i,j = 1,\ldots,m$

$$||vec(\Delta Y C) - (C \Delta X' \otimes I_m) vec(B)||^2_{I_T \otimes \Sigma_u^{-1}} + \sum_{k=1}^{P} \sum_{i,j=1}^{m} \frac{\lambda_{i,j,T}^{lag,k}}{|\check{B}_k(i,j)|^\gamma} |B_k(i,j)| \tag{17}$$

for fixed tuning parameters $\lambda_{i,j,T}^{lag,k}, \gamma$, where γ here and in the rank selection (16) might differ. Moreover, the pre-estimate $\check{B}$ in the adaptive Lasso weight can be taken from the partial least squares estimate $\widetilde{B} = (\Delta Y C \Delta X')(\Delta X C \Delta X')^{-1}$ 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

$$\begin{aligned}\widetilde{B}^R &= \arg\min \|vec(\Delta Y C) - (C\Delta X' \otimes I_m)vec(B)\|^2 \\ &\quad + \nu_T \textstyle\sum_{k=1}^{P} \sum_{i,j=1}^{m} |B_k(i,j)|^2\end{aligned} \tag{18}$$

The following Theorem 3.1 shows that this pre-estimate is consistent for appropriate choices of tuning parameters.

Theorem 3.1. *If the tuning parameter ν_T in the ridge regression (18) satisfies $\frac{\nu_T}{\sqrt{T}} \to_p 0$, then $\sqrt{T}(\widetilde{B}^R - B) = O_p(1)$ under Assumptions 2.1 and 3.1.*

The tuning parameter ν_T is designed for ridge regression only, and therefore independent of the rest of the paper. Such choice of tuning parameters has some important implications, in small sample it will mitigate the multi-collinearity, and in large sample it will achieve consistency.

As in the case of rank selection, a lag k should be included into the model, whenever $\hat{B}_k$ from the Lasso selection (17) 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 (17) as $\hat{p} = \max_{1 \leq k \leq P}\{k|\hat{B}_k \neq 0\}$.

Note that the residual transformation C in the lag selection criterion (17) 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. Generally, the proposed Ridge regression pre-step can potentially be further refined, e.g. by 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. We expect effects on the overall selection consistency results, however, to be only minor. Morevover, our separate two-step approach for rank and lag length can help alleviate the numerical instability caused by multi-collinearity in the lag selection step. 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 3.3.1 below).

3.2 Model selection consistency

This section states the asymptotic properties of the adaptive Lasso-VECM procedure. First, we show the result for the cointegrating rank selection according to criterion (16) which uses the residual transformation M in order to focus on the respective partial effect.

Theorem 3.2. *Suppose that* $\lambda_{i,j,T}^{rank}/\sqrt{T} \to 0$ *and* $T^{\frac{1}{2}(\gamma-1)}\lambda_{i,j,T}^{rank} \to \infty$. *Under Assumptions 2.1 and 3.1 the objective function (16) 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* (16).
2. $\sqrt{T}vec(\hat{R}_T' - R')_{\mathcal{A}} \to_d \mathrm{N}(0, (\Sigma_{z1z1.\Delta x} \otimes \Sigma_u^{-1})_{\mathcal{A}}^{-1}(\Sigma_{z1z1.\Delta x} \otimes \Sigma_u^{-1})_{\mathcal{A}}(\Sigma_{z1z1.\Delta x} \otimes \Sigma_u^{-1})_{\mathcal{A}}^{-1})$ *for* $r > 0$.

Thus Theorem 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 3.3.1 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 (17) 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 3.3. *Suppose that* $\lambda_{i,j,T}^{lag,k}/\sqrt{T} \to 0$ *and* $T^{\frac{1}{2}(\gamma-1)}\lambda_{i,j,T}^{lag,k} \to \infty$. *Under Assumptions 2.1 and 3.1 the objective function (17) 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* (17)
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 component covariance matrices defined in* (13).

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 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.

3.3 Important Refinements and Generalizations

3.3.1 Refined model estimation in higher dimensions

In this section, we show strategies for refined estimation and its corresponding asymptotic results when the error terms are weakly dependent. With our proposed adaptive Lasso techniques, we can select the true model with probability one for sufficiently many observations. Although both model selection criteria (16) and (17) 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

$$\begin{aligned} &\sum_{t=1}^{T} ||\Delta Y_t - \alpha \tilde{\beta}^{\dagger\prime} Y_{t-1} - \sum_{k=1}^{p} B_k \Delta Y_{t-k}||^2_{\Sigma_u^{-1}} \\ &+ \sum_{i=1}^{m}\sum_{j=1}^{r} \lambda^{rank}_{i,j,T} |\alpha(i,j)| + \sum_{k=1}^{p}\sum_{i,j=1}^{m} \lambda^{lag,k}_{i,j,T} |B_k(i,j)| \end{aligned} \tag{19}$$

where $\lambda^{rank}_{i,j,T}, \lambda^{lag,k}_{i,j,T}$ are tuning parameters. For no penalty $\lambda^{rank}_{i,j,T} = \lambda^{lag,k}_{i,j,T} = 0$, we recover the reduced rank regression estimates for α and B^p from (19).

We show that with appropriate choices of tuning parameters, the penalized estimates from (19) 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 5.1 will confirm, their finite-sample performance, however, is superior in particular for estimation but also for

prediction.

Theorem 3.4. *Denote* $B^p = [B_1, \ldots, B_p]$. *If* $\lambda^{rank}_{i,j,T}/\sqrt{T} \to_p 0$ *and* $\lambda^{lag,k}_{i,j,T}/\sqrt{T} \to_p 0$, *then the solution to problem (19) under Assumptions 2.1 and 3.1 satisfies:*

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

where $\Gamma^p_t = [Y'_{t-1}\beta, \Delta Y'_{t-1}, \ldots, \Delta Y'_{t-p}]'$ *and* $\frac{1}{T}\sum_{t=1}^{T} \Gamma^p_t \Gamma^{p\prime}_t \to_p \Sigma_{\Gamma^p\Gamma^p}$.

Theorem 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 3.2 and 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 5.1 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.

3.3.2 Model selection with dependent error terms

Here we illustrate how Assumption 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 3.2. *In the VECM as* (1) *the error term can admit the following linear dependence structure*

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

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

Here we consider weaker dependence in the residual process, therefore stronger assumption on κ_j is required in Assumption 3.2, which is a sufficient condition for the absolute summability assumption on κ_j, i.e. $\sum_{j=0}^{\infty} \|\kappa_j\|_2 < \infty$.

Lemma 3.2. *Under Assumption 3.2, the least squares estimate for* Π *in* (11) *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 A.3).*

The term Υ measures the correlation between u_t and $Z_{1,t-1}$ due to the auto-correlation of u_t under Assumption 3.2.

Define $\Xi = \begin{bmatrix} \beta' \\ \beta'_\perp \end{bmatrix}$, 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 3.1, we can conclude that

Lemma 3.3. *By the same notation as in Lemma 3.1 and under Assumption 3.2, the following results hold:*

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

Due to the bias term, we can't expect that the selection result 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 $i = 1, \ldots, m$

$$\sum_{t=1}^{T} \| \Delta Y_t - R'\widetilde{S}'Y_{t-1} \|_2^2 + \sum_{i=1}^{m} \frac{\lambda_{i,T}^{rank}}{||\widetilde{R}(i,)||_2^{\gamma}} ||R(i,)||_2 \tag{20}$$

Different from before, 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 3.3, 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 (20) in rank selection is expected. Besides, the first term in (20) 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 3.1.

Proposition 3.1. *Given Assumption 3.2, suppose that* $\lambda_{i,T}^{rank}$ *satisfies* $\frac{\lambda_{i,T}^{rank}}{\sqrt{T}} \to 0$ *and* $T^{\gamma-1}\lambda_{i,T}^{rank} \to \infty$*, the solution to* (20) *is consistent in selecting the right rank.*

When the dimension is higher, the variance of $\hat{R}$ from (16) generally increases due to the non-sparse structure within non-zero rows of $\hat{R}$.

4 Determination of network effects

In this section, we present the statistical properties for estimates of the impulse responses and the corresponding forecast error variance decomposition (FEVD) as building blocks for the network connectedness with an underlying large-dimensional VECM dynamics (1).

As shown in Park and Phillips (1989) and Phillips (1998) the impulse response functions are then given by the elements of the sequence of matrices Φ_j or certain linear combinations of the components of Φ_j in the MA-representation (5), depending on the information set containing the ordering of the shocks or structural relations among them. We get an estimate $\hat{\Phi}_j$ of Φ_h in (5) from estimates $\hat{A}$ of the coefficient matrices A as defined in (4) with $\hat{\Phi}_h = \hat{A}_{11}^h$ where $\hat{A}_{11}$ where is the upper left-hand $(m \times m)$ block of $\hat{A}$. The estimate $\hat{A}$ is obtained from the adaptive lasso procedures (16) and (17) that yield estimates for the components $\alpha\beta'$ and B_1 of A_{11} in (4). Alternatively, we can also use the refined two-step version (19) for the components.

Thus the following result holds for each component h of the impulse response function

Theorem 4.1. *Under Assumptions 2.1 and 3.1, let estimates* $\hat{\Phi}_j$ *of the impulse response matrices* Φ_j *be constructed such that all conditions for Theorem 3.2 and 3.3 are met. Then we get for each integer* $j \geq 0$*:*

$$\|\hat{\Phi}_j - \Phi_j\|_2 = O_P(\frac{1}{\sqrt{T}}) \ .$$

The above theorem shows that the MA coefficient matrices are $\sqrt{T}$-consistent. The respective rate corresponds to the usual stationary rate as expected given the definition of A. Note that the result directly generalizes to the case when $\hat{\Phi}_j$ are obtained from the refined two-step procedure (19) when the conditions of Theorem 3.4 are met. Moreover,

general time-dependent innovations as in Assumption 3.2 are admissible if estimates $\hat{\Phi}_j$ are produced from (20) and the conditions of Proposition 3.1 hold.

With the T-consistent standard least-squares estimate of Σ_u, Theorem 4.1 also implies consistency of the estimates of the impulse response components $IRF(j,h)$ by standard arguments.

From the MA representation (5), the forecast error of the optimal h-step ahead predictor $Y_{t,h}$ is $Y_{t+h} - Y_t = \sum_{j=0}^{h-1} \Phi_j u_{t+h-j}$. Its variance matrix, the h-step ahead forecast-error variance F_h is then

$$F_h \;=\; E(Y_{t+h} - Y_t)(Y_{t+h} - Y_t)' = \sum_{j=0}^{h-1} \Phi_j \Sigma_u \Phi_j' \; . \tag{21}$$

We employ a consistent plug-in estimator for all components in (21) in order to derive an estimate $\hat{F}_h$ for F_h.

Theorem 4.2. *Under the Assumptions of Theorem 4.1 we get for each h, we get*

$$\|\hat{F}_h - F_h\|_2 = O_P(\frac{1}{T}) \; .$$

The above result shows that the estimated forecast error variance matrices for finite forecast horizon h are T-consistent. Thus standard results imply $\sqrt{T}$-consistency for each forecast error variance decomposition (6) and thus consistency for all estimated network links based on the connectedness measures derived from (7).

5 Simulations and empirical findings

5.1 Simulations

In this section, we investigate the finite-sample performance of the proposed model selection methodology. We first study the estimation and prediction performance of our refined Lasso estimates in comparison to reduced rank method, this includes standard settings of dimension three for comparison with existing low dimensional techniques. Then 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 existing literature.

The results presented in this section are based on independent multivariate Gaussian innovations with covariance matrix $\Sigma_u = [\rho^{|i-j|}]_{i,j=1}^m$ for two cases of $\rho = 0.0$ and $\rho = 0.6$. Thus

our specifications also include cases of strong cross-sectional dependence. For example, the chosen vanishing pattern of correlations may correspond to increasing geographical distance in the case of the FX application presented in Section 5.2. For these settings, we use the general FGLS-type empirical versions of the objective functions (16) and (17) 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 1 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 focus on the penalties λ and γ for the rank selection.[4] Although lag and rank selections work independently, we find that choosing p first according to Theorem 3.3 leads to superior finite-sample choices of p which can then be used in setting P for numerically efficient rank selection in (16). 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 median values which actually hardly vary over all simulation runs. 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.

In this paper, we consider the following settings:

[4]For the lag selection, we chose the parameters identical to the rank selection. In practice, this could be further refined with different tuning parameters for each criterion, where the choice in the rank criterion is key as dealing with the nonstationary setting while the selection in the lag case is more robust as comparable to the standard stationary case.

$$
\begin{array}{lllll}
\text{model 1:} & m = 3 & r = 2 & p = 1 \\
\text{model 2:} & m = 8 & r = 4 & p = 1 \\
\text{model 3:} & m = 8 & r = 2 & p = 2 \\
\text{model 4:} & m = 16 & r = 8 & p = 1
\end{array}
$$

Model 1 For this standard three dimensional case, we choose a setting considered in Chao and Phillips (1999) for comparison purposes. The experiments 7 and 8 in Chao and Phillips (1999) are a trivariate VAR with one lag and two cointegration vectors entering a single equation of the system. In their setting, the Monte Carlo study has demonstrated that their criterion performs well in small samples. In addition to $\rho = 0.0$, we allow for strong cross-sectional dependence by choosing $\rho = 0.6$. Our rank and lag selection results indicate that lag selection performs well independent of the exact choice of tuning parameters with almost perfect results. More details are available in the Appendix C.

For the cases of 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 3.1. In particular, all unknown elements are drawn independently from $U[-1.5, 1.5]$. Therefore in the following settings (Model 2, 3 and 4), the model specifications are randomly chosen, see Appendix C for more details.

Model 2 and Model 3 These two models are both of dimension $m = 8$, where traditional methods cannot be employed either due to inconsistency in theory or because of numerical inefficiency. Note that for both model 2 and model 3, the results are based on a ridge regression pre-estimate (18) for the lag selection criterion (17) in order to handle multicollinearity effects. Lag selection results based on adaptive weights from least squares pre-estimates perform substantially inferior.[5]

The selection results for model 2 with $p = 1$ and $r = 4$ are represented in upper panel of Table 1. Note that the lag and rank selections work independently. In the table we report two values in each cell (the absolute numbers of correct rank/lag selections) using the same tuning parameters. If we take the values reported in different cells, we can easily compare the rank and lag selection results with different tuning parameters. In general, the results demonstrate perfect performance in rank and lag selection for a wide range of tuning parameters when $c \geq 1$ and $\gamma \geq 3$. This also holds even for the most difficult case: $\rho = 0.6$ and $T = 200$, while for all other settings the range of acceptable parameters is even wider. In comparison to the low-dimensional model 1, larger tuning parameters are preferred both for rank and lag selection due to the higher complexity of the true model.

[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/<u>97</u>	100/100	100/100
$\gamma = 4.0$	100/100	100/100	100/100
$\gamma = 5.0$	<u>100</u>/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/<u>90</u>
$\gamma = 3.0$	100/100	100/100	100/100
$\gamma = 4.0$	100/100	100/100	100/100
$\gamma = 5.0$	<u>100</u>/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/<u>88</u>	100/99	98/99
$\gamma = 4.0$	100/100	99/100	99/100
$\gamma = 5.0$	<u>100</u>/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/<u>88</u>	100/99	100/100
$\gamma = 4.0$	100/100	100/100	100/100
$\gamma = 5.0$	<u>100</u>/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/<u>91</u>	95/98	100/99
$\gamma = 3.0$	100/100	100/100	100/100
$\gamma = 4.0$	100/94	100/65	100/41
$\gamma = 5.0$	<u>100</u>/41	100/11	100/1

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

	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	<u>100</u>/<u>100</u>	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/<u>63</u>	80/80	90/92
$\gamma = 3.0$	92/100	97/99	99/97
$\gamma = 4.0$	98/90	99/48	98/17
$\gamma = 5.0$	<u>99</u>/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/<u>69</u>	100/85	100/94
$\gamma = 3.0$	100/100	100/100	100/100
$\gamma = 4.0$	100/100	100/100	100/100
$\gamma = 5.0$	<u>100</u>/99	100/56	100/26

Table 1: Each cell reports two values (the absolute numbers of correct rank/lag selections) by solving (16) and (17) for $b = 100$ repetitions for model 2 and 3 with $m = 8$, $r = 2$, $p = 2$. To compare the rank and lag selections with different tuning parameters, we can take the corresponding values reported in different cells. Underlining marks the choice with tuning parameters selected according to median BIC.

For model 3, the larger lag length $p = 2$ poses challenge in estimating the results. The selection of the tuning parameter $\gamma = 3.0$ results in very high correct estimates of both lag and rank selection results. In particular, for the case of 200 observations, larger tuning parameters are preferred for rank selection.

Model 4 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 (17). Generally, the simulation results show that lag selection works 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.

Model 4 ($T = 500$, $\rho = 0.0$)				**Model 4** ($T = 500$, $\rho = 0.6$)			
	$c = 1$	$c = 2$	$c = 3$		$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	<u>69</u>/<u>98</u>	98/100	100/100	$\gamma = 2.0$	11/<u>93</u>	58/100	84/100
$\gamma = 3.0$	100/100	78/100	46/100	$\gamma = 3.0$	<u>100</u>/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: Absolute numbers of correct rank/lag selections by solving (16) and (17) for $b = 100$ repetitions for model 4 with $m = 16,\ r = 8,\ p = 1$. Reporting style is as in Table 1.

For known true model specifications, we estimate all four models above according to the refined Lasso procedure (19) 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 (16) and $\widetilde{S}_r'$ consists of the first r rows of the orthonormal matrix defined in (15). 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 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 10). Similar results are reported in Table 11 for model 3 and Table 12 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 confirm the advantage of element-wise penalization on the loading matrix over penalization on 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. This might also drive the excellent forecasting performance in all considered model set-ups as the results in Tables 9-12 indicate.

5.2 Empirical results

There exists a sizable literature, such as e.g. Meese and Rogoff (1983) and Cheung et al. (2005), which concludes that a pure random walk model can hardly be beaten in forecasting floating exchange rates (FX) between countries by advanced time series methods. In particular, system information and cointegration structures could not be shown to yield any prediction advantages. Related work like Engel and West (2005) and Engel et al. (2015) among others, apply techniques such as panel data and factor model methods to predict exchange rates also for larger systems. In general, they obtain promising results which are, however, mixed with regard to beating the random walk benchmark. From economic theory, however, it is clear that system and equilibrium cointegration information on different underlying stochastic trends of exchange rates (see e.g. Baillie and Bollerslev (1989)) do carry valuable information which should provide performance gains in particular for longer horizons. Thus we use our proposed tailored lasso technique to estimate VECM for a moderate dimension portfolio consisting of 17 series. By exploiting the potential sparse cointegration relations within the FX market, we are able to also study the spillover network of the FX system gaining insights into important channels. Moreover, our method also provides improved forecasts without inclusion of additional information, where the benchmark is the random walk without

drift.

Our empirical analysis uses quarterly data from Engel et al. (2015).[6] We consider bilateral exchange rates y_{it} calculated as the end of quarter t logarithmic exchange rate of country i against the U.S. dollar (USD). We study 17 OECD countries: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Japan, Italy, Korea, Netherlands, Norway, Spain, Sweden, Switzerland, and the United Kingdom. For comparability with the original results of Engel et al. (2015) we use the same estimation period running from the first quarter of 1973 to fourth quarter of 2007 for a total of 140 observations.

	ADF, y_{it}	ADF, Δy_{it}	KPSS, y_{it}	KPSS, Δy_{it}
Australia	0.95	0.01	0.01	0.08
Austria	0.98	0.04	0.01	0.03
Belgium	0.50	0.01	0.10	0.10
Canada	0.39	0.01	0.01	0.10
Denmark	0.38	0.01	0.01	0.10
Finland	0.42	0.01	0.01	0.10
France	0.88	0.01	0.01	0.10
Germany	0.76	0.01	0.01	0.10
Japan	0.07	0.01	0.01	0.10
Italy	0.24	0.01	0.01	0.10
Korea	0.34	0.01	0.05	0.10
Netherlands	0.57	0.01	0.03	0.10
Norway	0.22	0.01	0.01	0.10
Spain	0.69	0.01	0.01	0.09
Sweden	0.71	0.01	0.01	0.05
Switzerland	0.50	0.01	0.01	0.10
United Kingdom	0.29	0.01	0.01	0.10

Table 3: The p-values for the panel unit root tests of FX time series for each country. For country i, y_{it} is the log value of FX, and Δy_{it} is the log return of FX. The ADF (Augmented Dickey-Fuller test) tests the null hypothesis that a unit root is present in a time series sample, and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) is used for testing a null hypothesis that an observable time series is stationary.

As a pre-check for the existence of cointegration relations, we apply panel unit root tests to both original y_{it} and differenced data Δy_{it}, with the corresponding p-values reported in Table 3. The results of ADF and KPSS tests indicate clearly that the differenced data Δy_{it} are stationary while y_{it} are not. This presence of unit roots in FX-rates was also documented in e.g. Baillie and Bollerslev (1989) and Diebold et al. (1994) where low-dimensional subsystems of cointegration were studied.

We start with the general VECM (1) by conducting both rank and lag estimation procedure. The simulation results in the previous section indicate robust model performance

[6]For a detailed description of the data and their sources we refer to Engel et al. (2015).

for a wide range of tuning parameters, where $\gamma = 3$ and a BIC-based choice of λ in both cases generally yielded convincing results. We follow this best-practice guidance in determining the tuning parameters and also set the upper bound for the lag selection as $P = 5$. Note that the estimation results are based on a ridge regression pre-estimate (18) for $\check{B}$ in (17) in order to handle multicollinearity effects, with the optimal tuning parameters selected by BIC. Then we obtain lag length $\hat{p} = 0$, and a cointegration rank of $rank(\hat{R}) = \hat{r} = 2$. Therefore the resulting model is as follows,

$$\Delta Y_t \quad = \quad \Pi Y_{t-1} + u_t \tag{22}$$

where Y_t is the vector composed of the stacked cross-sectional observations y_{it}, $i = 1, \ldots, 17$. We depict the time evolution of the two resulting cointegration factors in Figure 5 in the Appendix.

In the following, we illustrate the finite sample prediction performance gain from the proposed VECM in comparison to standard benchmarks in the last subsection. We also study how this can be used to study the network spillover effects in connectedness among FX rates using the network measures based on (7). As a benchmark for the determined VECM specification (22), we employ connectedness-based networks obtained from the directly estimated corresponding VAR(p) model in differences. This is of independent interest as such models have been widely used in the applied literature.

5.2.1 Static network analysis

We further construct the DY-network by computing variance decompositions and corresponding connectedness measures at horizon $H = 10$.[7] The graph of our full-sample FX market network defined in (6) is depicted in Figure 1. We observe a cluster of six closely interconnected European countries (France, Germany, Spain, Italy, Finland and the Netherlands) as highlighted by stronger color intensity in this graph, which are now part of the European monetary union (EMU). This can be explained by the economic integration among these countries, which involves the coordination of economic and fiscal policies, a common monetary policy, and a common currency, the euro among these Eurozone nations. Based on Figure 1, the network graph in Figure 2 highlights the significant pairwise directional connectedness among the six EMU countries, in particular for the countries of France, Germany, Italy, Finland and the Netherlands.

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,

[7]Presented results are robust for H in the range of 8-12.

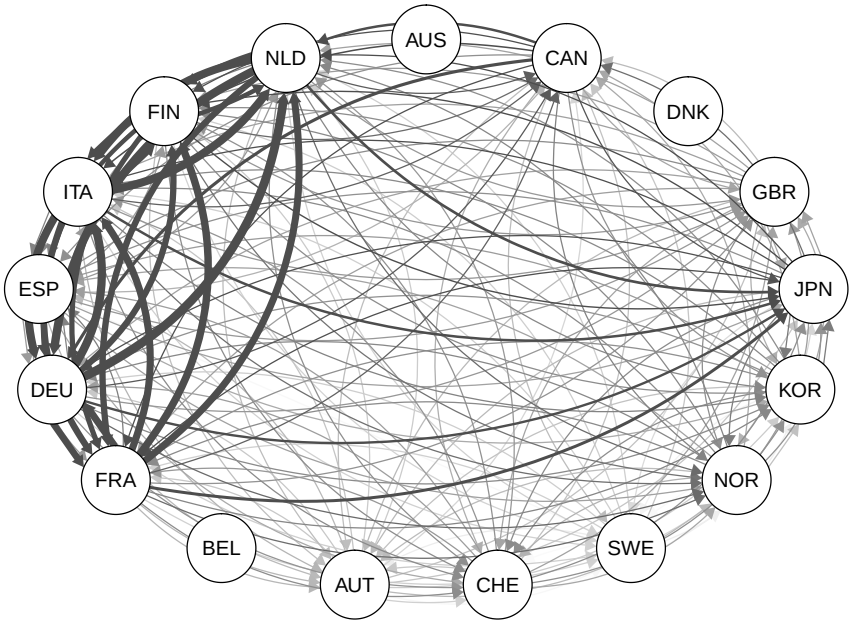

Figure 1: The graph for full-sample FX market network for 17 OECD countries based on the estimation of our VECM model (22). In this graph, we select the 90th quantile to cut the scaling of edges in width and color saturation. Edges with absolute weights over this value will have the strongest color intensity and become wider the stronger they are, and edges with absolute weights under this value will have the smallest width and become vaguer the weaker the weight (see Epskamp et al. (2012)).

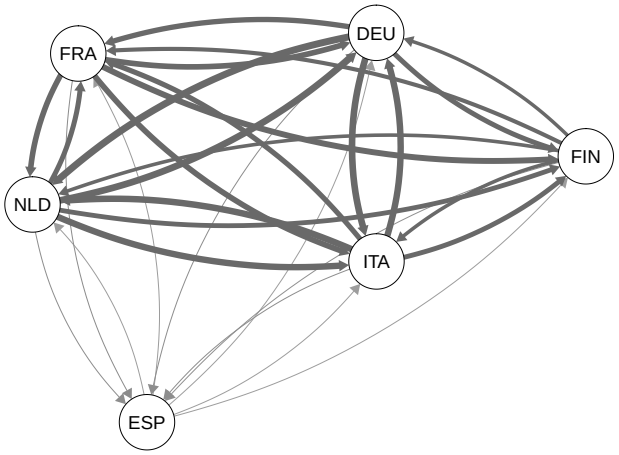

Figure 2: The graph for the FX market network among several European countries. As before, we select the 90th quantile to cut the scaling of edges in width and color saturation.

- 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.

- 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.

- 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.

Table 4 reports the above four centrality measures and three connectedness measures defined in (7) and (8) for all the sample countries. We observe higher negative net connectedness for France, Germany, Italy, Finland and Netherland, indicating a net connectedness receiver behavior. Besides, Denmark and Sweden have relatively larger positive levels of net connectedness making these two non-EMU countries net connectedness transmitter for the FX market. Note that the obtained above results for the different network measures are stable across the a wide range of tuning parameter choices as indicated in Figures 6, 7, and 8 in the Appendix.

For the estimation of the benchmark VAR specification in differences for large dimensions we use VAR combined with Lasso. A VAR model with lag length of one was selected by BIC. The resulting DY-network connectedness and centrality measures are reported in Table 5. The graph for the full-sample FX market network is shown in Figure 3. The topology of the graph and the observed spillover effects differ substantially from the VECM results which are generally known to have higher finite sample accuracy (see e.g. Engle and Yoo (1987)). The estimation results based on the benchmark VAR-FEVD are mixed. We also observe strong pairwise connectedness among several European countries which are also net connectedness receivers, but the components are Denmark, Netherland, Germany, France, Belgium and Austria. Among them, Denmark is not part of the EMU and retains its own monetary policy and currency. In addition, two non-European countries Australia and Japan become net connectedness transmitters.

In the closing part of this section we show that our VECM based connectedness is not sensitive to parameter choices. The results presented in Figure 2 show the network topology

	$C^H_{from,i}$	$C^H_{to,i}$	$C^H_{net,i}$	$indeg(i)$	$outdeg(i)$	$Bet(i)$	$Clos(i)$
Australia	0.97	0.45	0.52	0.01	0.01	0.00	0.02
Canada	5.23	5.33	-0.10	0.45	0.52	16.00	0.07
Denmark	3.05	1.71	1.34	0.05	0.05	29.00	0.02
UK	5.15	4.78	0.37	0.39	0.38	0.00	0.06
Japan	5.30	6.17	-0.87	0.52	0.52	0.00	0.07
Korea	5.17	4.94	0.23	0.41	0.41	0.00	0.07
Norway	5.23	5.44	-0.21	0.46	0.40	26.00	0.07
Sweden	4.66	2.81	1.85	0.22	0.23	40.00	0.06
Switzerland	5.15	4.76	0.39	0.40	0.39	0.00	0.06
Austria	5.00	3.85	1.15	0.32	0.29	0.00	0.06
Belgium	1.80	0.95	0.85	0.02	0.01	0.00	0.01
France	5.32	6.38	-1.06	0.53	0.56	0.00	0.07
Germany	5.32	6.40	-1.08	0.54	0.56	0.00	0.07
Spain	5.20	5.26	-0.06	0.43	0.38	0.00	0.07
Italy	5.33	6.43	-1.10	0.54	0.56	2.00	0.07
Finland	5.32	6.43	-1.11	0.54	0.54	2.00	0.07
Netherlands	5.33	6.43	-1.10	0.54	0.56	0.00	0.07

Table 4: The "from", "to" and "net" connectedness for the sample countries based on the estimation of our VECM model (22).

	$C^H_{from,i}$	$C^H_{to,i}$	$C^H_{net,i}$	$indeg(i)$	$outdeg(i)$	$Bet(i)$	$Clos(i)$
Australia	3.27	1.75	1.52	0.06	0.06	43.00	0.03
Canada	2.29	1.05	1.24	0.03	0.02	0.00	0.02
Denmark	5.34	6.41	-1.07	0.54	0.54	3.00	0.07
UK	5.03	3.94	1.09	0.32	0.28	0.00	0.06
Japan	4.77	2.89	1.88	0.24	0.25	15.00	0.06
Korea	0.77	0.56	0.21	0.01	0.01	0.00	0.01
Norway	5.25	5.54	-0.29	0.46	0.40	33.00	0.07
Sweden	5.18	4.93	0.25	0.40	0.39	0.00	0.07
Switzerland	5.24	5.23	0.01	0.45	0.53	8.00	0.07
Austria	5.33	6.35	-1.02	0.54	0.56	0.00	0.07
Belgium	5.33	6.35	-1.02	0.54	0.57	0.00	0.07
France	5.32	6.22	-0.90	0.52	0.52	0.00	0.07
Germany	5.33	6.32	-0.99	0.54	0.56	0.00	0.07
Spain	5.19	4.97	0.22	0.41	0.39	0.00	0.06
Italy	5.22	5.14	0.08	0.43	0.43	0.00	0.07
Finland	5.24	5.42	-0.18	0.44	0.41	14.00	0.07
Netherlands	5.34	6.38	-1.04	0.54	0.56	0.00	0.07

Table 5: The "from", "to" and "net" connectedness for the sample countries based on VAR(1)-Lasso estimation.

when we consider alternative specifications.

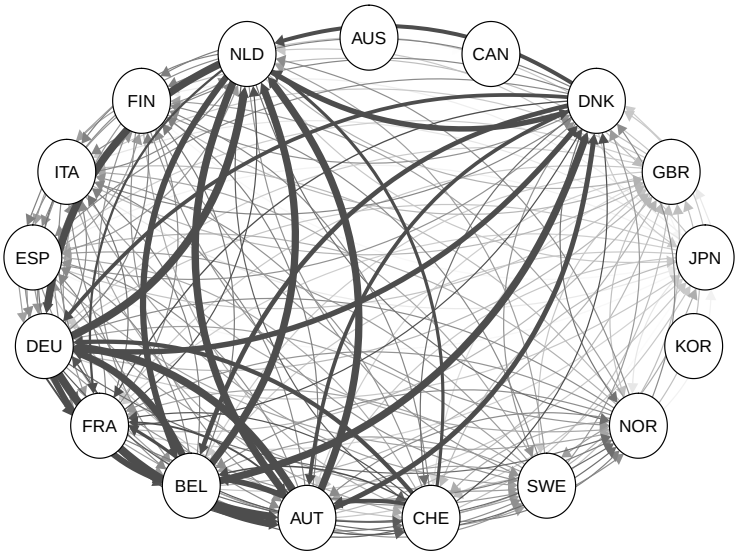

Figure 3: The graph for full-sample FX market network for 17 OECD countries, based on VAR(1)-Lasso estimation. We select the 90th quantile to cut the scaling of edges in width and color saturation.

5.2.2 Dynamic network analysis

We now study the dynamic network using rolling estimation, and compare the dynamic total connectedness from our VECM-FEVD with the one based on the VAR model estimated over the same rolling window. The number of observations used in the rolling sample to compute prediction is 120 or correspondingly thirty years, and we examine dynamic evolution of the network for the following five years (20 observations). In each window, we repeat model selection and conduct the proposed technique to obtain the sparse estimates.

We first calculate full sample system-wide connectedness for each window by summing up the total directional connectedness whether "from" or "to". In general, the full sample system-wide connectedness reflects the overall uncertainty that has arisen in the system. The dynamic pattern of the system-wide connectedness is shown in the left panel of Figure 4. The VECM based system-wide connectedness is larger than the VAR based system-wide connectedness. We interpret this result as the VECM based connectedness capturing the impact of long-run relationships that affected the FX market, particularly the EMU and non-EMU countries.

To assess the system-wide interaction, we further decompose the full sample system-wide connectedness into two parts: cross-EMU connectedness and within-EMU connectedness as shown in the right panel of Figure 4. In both cases the within-EMU spillovers generally exceed cross-EMU values at all time-points. The differences between the two parts are

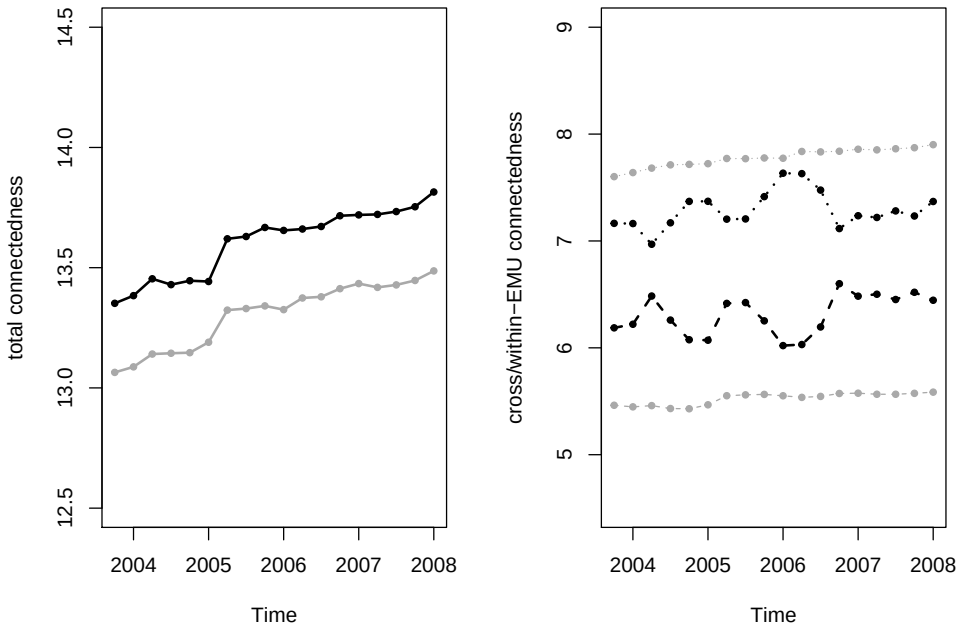

Figure 4: The time-varying network for the system-wide connectedness from April 2003 to January 2008, based on the the estimation of the VECM (black) and VAR (grey) models. The left panel is the time-varying full sample connectedness, it can be decomposed into two parts: the within-EMU connectedness (upper dotted curves in the right panel) and the respective cross-EMU connectedness (lower dashed curves in the right panel).

much more pronounced and smoothed over time in the VAR, while the VECM components indicate substantial variation in time in opposite directions. As a plausiblity check, we have compared the within-EMU-connectedness to the EUR/USD exchange rates and found an overall correlation of 0.315 for the VECM-based component dominating the VAR.

5.2.3 Out-of-sample forecasting performance

We compare the out-of-sample forecasting performance of our model (22) to two benchmark models, the simple random walk and VAR(1)-Lasso. We use the first 120 observations (i.e. 30 years) in our sample for estimation and the remaining 20 observations for evaluation.

We already have Y_t as the actual data series, let $\hat{Y}_{i,t}^h$ denote the ith competing h-step forecasting series and the out-of-sample forecasting errors from the ith competing models are defined as $\epsilon_{i,t}^h = Y_t - \hat{Y}_{i,t}^h$. In this paper, h is set to be 1, and the superscript h is omitted in the following context. Table 6 compares the mean μ and the corresponding 5% confidence interval, and standard deviation sd for the out-of-sample forecast errors

$\epsilon_{i,t}$ [8] of our model (left panel) and the random walk benchmark (middle panel) and the VAR benchmark (right panel). The results clearly show the superiority of our technique throughout all countries in the sample. We observe not only smaller values of forecast errors, but also narrower confidence intervals for our VECM model.

	$\mu(\epsilon_{VECM,t})$			$sd(\epsilon_{VECM,t})$	$\mu(\epsilon_{RW,t})$			$sd(\epsilon_{RW,t})$	$\mu(\epsilon_{VAR,t})$			$sd(\epsilon_{VAR,t})$
Australia	0.02	(-0.01	0.05)	0.01	0.17	(-0.11	0.48)	0.12	0.11	(-0.75	0.88)	0.45
Canada	0.10	(0.04	0.17)	0.03	0.14	(0.00	0.30)	0.05	0.15	(-0.36	0.71)	0.35
Denmark	0.07	(-0.03	0.16)	0.04	0.09	(-0.26	0.45)	0.27	0.09	(-0.65	0.77)	0.40
UK	0.05	(-0.07	0.19)	0.06	0.13	(-0.18	0.46)	0.19	0.06	(-0.42	0.68)	0.33
Japan	0.07	(-0.08	0.22)	0.07	-0.01	(-0.37	0.32)	0.41	0.02	(-0.47	0.71)	0.39
Korea	0.15	(-0.02	0.33)	0.08	0.22	(-0.15	0.57)	0.16	0.11	(-0.34	0.82)	0.31
Norway	0.09	(-0.09	0.29)	0.10	0.13	(-0.19	0.44)	0.20	0.10	(-0.59	0.83)	0.45
Sweden	0.02	(-0.19	0.23)	0.11	0.14	(-0.20	0.47)	0.21	0.05	(-0.92	0.77)	0.54
Switzerland	0.00	(-0.22	0.24)	0.12	0.02	(-0.36	0.42)	0.38	0.06	(-0.68	0.84)	0.41
Austria	-0.03	(-0.28	0.22)	0.13	0.06	(-0.30	0.42)	0.31	0.10	(-0.64	0.84)	0.43
Belgium	-0.01	(-0.27	0.26)	0.14	0.08	(-0.28	0.46)	0.30	0.10	(-0.64	0.84)	0.43
France	0.02	(-0.25	0.31)	0.14	0.11	(-0.24	0.47)	0.25	0.09	(-0.64	0.85)	0.40
Germany	0.01	(-0.29	0.31)	0.15	0.06	(-0.30	0.43)	0.32	0.11	(-0.64	0.84)	0.43
Spain	0.04	(-0.29	0.36)	0.16	0.15	(-0.19	0.48)	0.20	0.09	(-0.68	0.82)	0.43
Italy	0.04	(-0.30	0.38)	0.17	0.16	(-0.20	0.51)	0.21	0.06	(-0.67	0.80)	0.42
Finland	0.04	(-0.30	0.40)	0.17	0.11	(-0.21	0.43)	0.22	0.08	(-0.66	0.80)	0.42
Netherlands	0.08	(-0.27	0.45)	0.18	0.06	(-0.30	0.43)	0.31	0.10	(-0.64	0.84)	0.43

Table 6: Comparison of the out-of-sample forecast errors $\epsilon_{i,t}$, with our model on the left panel, the random walk benchmark in the middle and the VAR benchmark on the right panel.

We also compute the mean squared prediction error $MSE_{i,t} = \frac{1}{T}\sum_{t=1}^{T} \epsilon_{i,t}^2$ for each sample country, and apply a one sided hypothesis test on $H_0 : MSE_{VECM,t} \geq MSE_{RW,t}$ against $H_1 : MSE_{VECM,t} < MSE_{RW,t}$. The p-value of the t-test is 0.00026. We therefore reject the null hypothesis, indicating that there is strong evidence of smaller MSE for our technique.

In addition to the comparative MSE evaluation, we further use the Diebold and Mariano (DM) test (see Diebold and Mariano (2002) and Diebold (2015)) for comparing predictive accuracy. Denote the loss associated with forecast error $\epsilon_{i,t}$ by $L(\epsilon_{i,t})$; here we consider the squared-error (SE) loss function $L_1(\epsilon_{i,t}) = \sum_{t=1}^{T} \epsilon_{i,t}^2$ and the absolute-error (AE) loss function $L_2(\epsilon_{i,t}) = \sum_{t=1}^{T} |\epsilon_{i,t}|$. Table 7 shows that generally $VECM$ clearly outperforms the RW and VAR where the RW is the runner-up.

6 Conclusion

In this paper, we provide a novel technique for estimating large spillover networks of nonstationary systems in VECM framework. This elementwise Lasso-type technique does

[8] We use the standard bootstrap (Hubrich and West (2010) and Engel et al. (2015)) with 1000 rsepetitions for each point.

	$H_0 : E\{L(\epsilon_{VECM})\} < E\{L(\epsilon_{RW})\}$		$H_0 : E\{L(\epsilon_{VECM})\} < E\{L(\epsilon_{VAR})\}$		$H_0 : E\{L(\epsilon_{VAR})\} < E\{L(\epsilon_{RW})\}$	
	DM-AE	DM-SE	DM-AE	DM-SE	DM-AE	DM-SE
Australia	1.000	0.999	1.000	0.994	0.021	0.017
Canada	0.999	0.998	1.000	0.998	0.009	0.006
Denmark	0.999	0.994	1.000	0.998	0.000	0.003
UK	1.000	1.000	1.000	0.995	0.023	0.016
Japan	0.968	0.862	1.000	0.999	0.000	0.001
Korea	1.000	1.000	0.998	0.949	0.458	0.247
Norway	1.000	0.997	1.000	0.999	0.000	0.002
Sweden	1.000	0.999	1.000	0.999	0.000	0.001
Switzerland	0.010	0.005	1.000	0.995	0.000	0.004
Austria	0.905	0.897	1.000	0.998	0.000	0.002
Belgium	0.999	0.993	1.000	0.998	0.000	0.003
France	1.000	0.998	1.000	0.998	0.001	0.004
Germany	0.962	0.946	1.000	0.998	0.000	0.002
Spain	1.000	0.999	1.000	0.998	0.006	0.007
Italy	1.000	0.999	1.000	0.998	0.006	0.007
Finland	1.000	0.997	1.000	0.998	0.001	0.003
Netherlands	0.981	0.967	1.000	0.998	0.000	0.002

Table 7: The p-values for the Diebold-Mariano tests based on different models for each country. For country i, we compare the forecasting two models using both the Diebold-Mariano test by absolute-error loss (DM-AE) and the Diebold-Mariano test by squared-error loss (DM-SE).

not only determine cointegration rank and autoregressive lags of the large nonstationary system, but also allows to directly assess the non-zero elements in the cointegration vector, the resulting VECM estimation is then associated with network structure. The tailoring of the procedure to moderate large but fixed dimensions also keeps the technical prerequisites for statistical validity to the standard low dimensional assumptions, making the technique easily accessible for practitioners in most relevant application cases.

We report results on model selection consistency, derive the asymptotic distribution of estimates and propose refinements under general assumptions on the innovation. We also report the statistical properties for network estimation under standard assumptions. The excellent finite sample performance of the proposed technique is demonstrated in a comprehensive simulation study. In an application to a system of FX rates, we study the spillover effects in the FX market among 17 OECD countries.

References

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.

Anderson, T. (2002). Reduced rank regression in cointegrated models. *Journal of Econometrics*, 106(2):203–216.

Baillie, R. T. and Bollerslev, T. (1989). Common stochastic trends in a system of exchange rates. *The Journal of Finance*, 44(1):167–181.

Barigozzi, M. and Brownlees, C. (2019). Nets: Network estimation for time series. *Journal of Applied Econometrics*, 34(3):347–364.

Billio, M., Getmansky, M., Lo, A. W., and Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of financial economics*, 104(3):535–559.

Boswijk, H. P., Jansson, M., and Nielsen, M. Ø. (2015). Improved likelihood ratio tests for cointegration rank in the var model. *Journal of Econometrics*, 184(1):97–110.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

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.

Cheung, Y.-W., Chinn, M. D., and Pascual, A. G. (2005). Empirical exchange rate models of the nineties: Are any fit to survive? *Journal of international money and finance*, 24(7):1150–1175.

Dahlhaus, R. (2000). Graphical interaction models for multivariate time series. *Metrika*, 51(2):157–172.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests. *Journal of Business & Economic Statistics*, 33(1):1–1.

Diebold, F. X., Gardeazabal, J., and Yilmaz, K. (1994). On cointegration and exchange rate dynamics. *The Journal of Finance*, 49(2):727–735.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144.

Diebold, F. X. and Yilmaz, K. (2009). Measuring financial asset return and volatility spillovers, with application to global equity markets. *The Economic Journal*, 119(534):158–171.

Diebold, F. X. and Yilmaz, K. (2012). Better to give than to receive: Predictive directional measurement of volatility spillovers. *International Journal of Forecasting*, 28(1):57–66.

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.

Eichler, M. (2007). Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, 137(2):334–353.

Eichler, M. (2012). Causal inference in time series analysis. *Causality: Statistical Perspectives and Applications*, pages 327–354.

Engel, C., Mark, N. C., and West, K. D. (2015). Factor model forecasts of exchange rates. *Econometric Reviews*, 34(1-2):32–55.

Engel, C. and West, K. D. (2005). Exchange rates and fundamentals. *Journal of political Economy*, 113(3):485–517.

Engle, R. and Granger, C. (1987). Co-integration and error correction: represen- tation, estimation and testing. *Econometrica*, 55:257–276.

Engle, R. F. and Yoo, B. S. (1987). Forecasting and testing in co-integrated systems. *Journal of Econometrics*, 35(1):143 – 159.

Epskamp, S., Cramer, A. O., Waldorp, L. J., Schmittmann, V. D., Borsboom, D., et al. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of statistical software*, 48(4):1–18.

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.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239.

Hubrich, K., Lütkepohl, H., and Saikkonen, P. (2001). A review of systems cointegration tests. *Econometric Reviews*, 20(3):247–318.

Hubrich, K. and West, K. D. (2010). Forecast evaluation of small nested model sets. *Journal of Applied Econometrics*, 25(4):574–594.

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.

Kock, A. B. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2):325–344.

Koop, G., Pesaran, M. H., and Potter, S. M. (1996). Impulse response analysis in nonlinear multivariate models. *Journal of econometrics*, 74(1):119–147.

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.

Liang, C. and Schienle, M. (2019). Determination of vector error correction models in high dimensions. *Journal of Econometrics*, 208(2):418 – 441.

Liao, Z. and Phillips, P. C. (2015). Automated estimation of vector error correction models. *Econometric Theory*, 31(03):581–646.

Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Springer Publishing Company, Incorporated.

Medeiros, M. C. and Mendes, E. F. (2016). l1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors. *Journal of Econometrics*, 191(1):255–271.

Meese, R. A. and Rogoff, K. (1983). Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of international economics*, 14(1-2):3–24.

Onatski, A. and Wang, C. (2018). Alternative asymptotics for cointegration tests in large vars. *Econometrica*, 86(4):1465–1478.

Park, J. Y. and Phillips, P. C. (1989). Statistical inference in regressions with integrated processes: Part 2. *Econometric Theory*, 5(1):95–131.

Pesaran, H. H. and Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. *Economics letters*, 58(1):17–29.

Phillips, P. C. (1998). Impulse response and forecast error variance asymptotics in non-stationary vars. *Journal of econometrics*, 83(1-2):21–56.

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.

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.

Wilms, I. and Croux, C. (2016). Forecasting Using Sparse cointegration. *International Journal of Forecasting*, 32:12561267.

Xiao, Z. and Phillips, P. C. (1999). Efficient detrending in cointegrating regression. *Econometric Theory*, 15:519–548.

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, R., Robinson, P., and Yao, Q. (2018). Identifying cointegration by eigenanalysis. *Journal of the American Statistical Association*, 0(0):1–12.

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.

A Proofs

Lemma A.1. *Under Assumptions 2.1 and 3.1, the partial least squares estimate* $\widetilde{\Pi}$ *defined in (14) satisfies*

$$\begin{aligned} & vec[Q(\widetilde{\Pi}-\Pi)Q^{-1}D_T] \\ \to_d & \begin{bmatrix} N(0,\Sigma_{z1z1.\Delta x}^{-1}\otimes\Sigma_v) \\ vec\Big\{\Sigma_v^{1/2}(\int_0^1 W_{m-r}^{\dagger}dW_m')'(\int_0^1 W_{m-r}^{\dagger}W_{m-r}^{\dagger\prime}ds)^{-1}(\alpha_\perp'\Sigma_u\alpha_\perp)^{-\frac{1}{2}}\Theta_{22}^{-1} \end{bmatrix} \end{aligned}$$

where $D_T = diag(\sqrt{T}I_r, TI_{m-r})$, $\Sigma_v = Q\Sigma_u Q'$, $Z_{-1} = \beta' Y_{-1}$, $\frac{1}{T}Z_{-1}MZ_{-1}' \to_p \Sigma_{z1z1.\Delta x} = \Sigma_{z1z1} - \Sigma_{z1\Delta x}\Sigma_{\Delta x\Delta x}^{-1}\Sigma_{\Delta xz1}$ *with all the component covariance matrices defined in* (13)*;* $W_{m-r}^{\dagger} = (\alpha_\perp'\Sigma_u\alpha_\perp)^{-\frac{1}{2}}[0_{(m-r)\times r}, I_{m-r}]\Sigma_v^{\frac{1}{2}}W_m$*, and* $W_{m-r}^{\dagger}, W_m$ *are standard Brownian motions with dimension* $m-r, m$ *respectively and the exact from of* Θ *is defined as* (23) *and* (24) *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 A.2. *With the notation defined in Section 3.1, we have*

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

where $\Sigma_{\Delta x\Delta x.z1} = \Sigma_{\Delta x\Delta x} - \Sigma_{\Delta xz1}\Sigma_{z1z1}^{-1}\Sigma_{z1\Delta x}$.

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

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.

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

□

Proof for Lemma A.1

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

$$\Delta Z_t = Q\Pi Q^{-1} Z_{t-1} + \psi_t \tag{23}$$

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

$$\psi_t = \Theta(L)v_t \tag{24}$$

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), $\frac{1}{T}\Delta XZ_{2,-1}' = O_p(1)$, $\frac{1}{T}Z_{1,-1}\Delta X' = O_p(1)$ 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.

$$\begin{aligned}\frac{1}{T^2}Z_{2,-1}MZ_{2,-1}' &= \frac{1}{T^2}Z_{2,-1}Z_{2,-1}' - \frac{1}{T}\frac{1}{T}Z_{2,-1}\Delta X'(\frac{1}{T}\Delta X\Delta X)^{-1}\frac{1}{T}\Delta XZ_{2,-1}' \\ &\to_d \Theta_{22}(\alpha_\perp'\Sigma_u\alpha_\perp)^{1/2}\int_0^1 W_{m-r}(s)W_{m-r}'(s)ds(\alpha_\perp'\Sigma_u\alpha_\perp)^{1/2}\Theta_{22}'\end{aligned}$$

3. Distribution of $QUMZ_{-1}'D_T^{-1}$*:*

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

The last equality follows from $\frac{1}{T^{\frac{3}{2}}}\Delta XZ_{2,-1} \to_p 0$ as shown in Ahn and Reinsel (1990). Since we have shown that $\frac{1}{T}Z_{1,-1}MZ'_{1,-1} \to_p \Sigma_{z1z1.\Delta x}$, $\frac{1}{\sqrt{T}}vec(VMZ_{1,-1}) \to_d N(0, \Sigma_{z1z1.\Delta x} \otimes \Sigma_v)$. 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 3.1

The proof directly follows from Lemma A.1. □

Proof of Theorem 3.1

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 $-(\frac{1}{n}X'X + \frac{\nu}{n}I_p)^{-1}\frac{\nu}{\sqrt{n}}\beta \to_p 0$ if $\frac{\nu}{\sqrt{n}} \to_p 0$. Therefore $\lim_{T\to\infty}\widetilde{B}_R = B$ holds. □

Proof of Theorem 3.2

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 $\Delta_T(E_R) = \Psi_T(E_R) - \Psi_T(0)$.

$$\begin{aligned}\Delta_T(E_R) &= vec(E_R D_T^{-1})'(\widetilde{S}' Y_{-1} \otimes I_m)(I_T \otimes \Sigma_u^{-1})(Y_{-1}'\widetilde{S} \otimes I_m) vec(E_R D_T^{-1}) \\ &- 2vec(U)'(I_T \otimes \Sigma_u^{-1})(Y_{-1}'\widetilde{S} \otimes I_m) vec(E_R D_T^{-1}) \\ &+ \sum_{i,j=1}^{m} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^\gamma} (|R(i,j) + E_R D_T^{-1}(i,j)| - |R(i,j)|) \\ &= vec(E_R)'(D_T^{-1}\widetilde{S}' Y_{-1} \otimes I_m)(I_T \otimes \Sigma_u^{-1})(Y_{-1}'\widetilde{S} D_T^{-1} \otimes I_m) vec(E_R) \\ &- 2vec(\Sigma_u^{-1} U Y_{-1}'\widetilde{S} D_T^{-1})' vec(E_R) \\ &+ \sum_{i,j=1}^{m} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^\gamma} (|R(i,j) + E_R D_T^{-1}(i,j)| - |R(i,j)|) \qquad (25) \\ &= vec(E_R)'(D_T^{-1}\widetilde{S}' \sum_{t=1}^{T} Y_{t-1} Y_{t-1}'\widetilde{S} D_T^{-1} \otimes \Sigma_u^{-1}) vec(E_R) \\ &- 2vec(\sum_{t=1}^{T} \Sigma_u^{-1} u_t Y_{t-1}'\widetilde{S} D_T^{-1})' vec(E_R) \\ &+ \sum_{i,j=1}^{m} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^\gamma} (|R(i,j) + E_R D_T^{-1}(i,j)| - |R(i,j)|)\end{aligned}$$

In Lemma 3.1 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: $0 < r < m$

$$\sum_{t=1}^{T} Y_{t-1}Y_{t-1}' = Q^{-1}D_T D_T^{-1}\sum_{t=1}^{T} Z_{t-1}Z_{t-1}' D_T^{-1}D_T Q'^{-1}$$

$$= Q^{-1}D_T \begin{pmatrix} T^{-1}\sum_{t=1}^{T} Z_{1,t-1}Z_{1,t-1}' & T^{-3/2}\sum_{t=1}^{T} Z_{1,t-1}Z_{2,t-1}' \\ T^{-3/2}\sum_{t=1}^{T} Z_{2,t-1}Z_{1,t-1}' & T^{-2}\sum_{t=1}^{T} Z_{2,t-1}Z_{2,t-1}' \end{pmatrix} D_T Q'^{-1}$$

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

$$\begin{aligned} & D_T^{-1}\widetilde{S}'\sum_{t=1}^{T} Y_{t-1}Y_{t-1}'\widetilde{S}D_T^{-1} \\ = & \begin{bmatrix} I_r + O_p(\frac{1}{T}) & \sqrt{T}O_p(\frac{1}{T}) \\ \frac{1}{\sqrt{T}}\widetilde{S}_2'q_1 & \widetilde{S}_2'q_2 \end{bmatrix} \begin{pmatrix} T^{-1}\sum_{t=1}^{T} Z_{1,t-1}Z_{1,t-1}' & T^{-3/2}\sum_{t=1}^{T} Z_{1,t-1}Z_{2,t-1}' \\ T^{-3/2}\sum_{t=1}^{T} Z_{2,t-1}Z_{1,t-1}' & T^{-2}\sum_{t=1}^{T} Z_{2,t-1}Z_{2,t-1}' \end{pmatrix} \\ & \begin{bmatrix} I_r + O_p(\frac{1}{T}) & \frac{1}{\sqrt{T}}q_1'\widetilde{S}_2 \\ \sqrt{T}O_p(\frac{1}{T}) & q_2'\widetilde{S}_2 \end{bmatrix} \qquad (26) \\ \to_d & \begin{bmatrix} \Sigma_{z1z1} & 0 \\ 0 & \widetilde{S}_2'q_2\left(\left([0 \quad I_{m-r}]\Sigma_v^{1/2}(\int_0^1 W_m W_m' ds)\Sigma_v^{1/2}\begin{bmatrix} 0 \\ I_{m-r} \end{bmatrix}\right)^{-1}\right)q_2'\widetilde{S}_2 \end{bmatrix} \end{aligned}$$

For the second term in equation (25), we have:

$$\begin{aligned} & vec(\Sigma_u^{-1}(\sum_{t=1}^{T} u_t Y_{t-1}')\widetilde{S}D_T^{-1}) = vec(\Sigma_u^{-1}(\sum_{t=1}^{T} u_t Y_{t-1}'Q'D_T^{-1})D_T Q'^{-1}\widetilde{S}D_T^{-1}) \\ = & vec(\begin{bmatrix} T^{-1/2}\sum \Sigma_u^{-1}u_t Z_{1,t-1}' & T^{-1}\sum \Sigma_u^{-1}u_t Z_{2,t-1}' \end{bmatrix} \begin{bmatrix} I_r + O_p(\frac{1}{T}) & \frac{1}{\sqrt{T}}q_1'\widetilde{S}_2 \\ \sqrt{T}O_p(\frac{1}{T}) & q_2'\widetilde{S}_2 \end{bmatrix}) \\ \to_d & \begin{bmatrix} N(0, \Sigma_{z1z1}\otimes\Sigma_u^{-1}) \\ vec\{\Sigma_u^{-1}Q^{-1}\Sigma_v^{\frac{1}{2}}(\int_0^1 W_m dW_m')'\Sigma_v^{\frac{1}{2}}\begin{bmatrix} 0 \\ I_{m-r} \end{bmatrix}q_2'\widetilde{S}_2\} \end{bmatrix} \qquad (27) \end{aligned}$$

Next we should pay attention to the last term in eq. (25).
For the first r columns of matrix R', the convergence rate of the least square estimator is $\sqrt{T}$. Therefore, if $R(i,j) \neq 0$, $\hat{w}_{i,j} = |\widetilde{R}(i,j)|^{-\gamma} \to_p |R(i,j)|^{-\gamma}$ and $\sqrt{T}(|R(i,j) + \frac{1}{\sqrt{T}}E_R(i,j)| - |R(i,j)|) \to sign(R(i,j))|E_R(i,j)|$. By Slutsky's theorem, we have $\frac{\lambda_{i,j,T}^{rank}}{\sqrt{T}}\hat{w}_{i,j}\sqrt{T}(|R(i,j) + \frac{1}{\sqrt{T}}E_R(i,j)| - |R(i,j)|) \to_p 0$.
If $R(i,j) = 0$, $T^{-\frac{\gamma}{2}}\hat{w}_{i,j} = O_p(1)$ and $\sqrt{T}(|R(i,j) + \frac{1}{\sqrt{T}}E_R(i,j)| - |R(i,j)|) \to |E_R(i,j)|$.
By Slutsky's theorem, we have $\frac{\lambda_{i,j,T}^{rank}T^{\frac{\gamma}{2}}}{\sqrt{T}}T^{-\frac{\gamma}{2}}\hat{w}_{i,j}\sqrt{T}(|R(i,j) + \frac{1}{\sqrt{T}}E_R(i,j)| - |R(i,j)|) \to_p \infty$.

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 $\frac{\lambda_{i,j,T}^{rank}}{T}T^{\gamma}|T\widetilde{R}(i,j)|^{-\gamma} \to_p \infty$, where $|T\widetilde{R}(i,j)| = O_p(1)$.

Thus, $\Delta_T(E_R) \to_d \Delta(E_R)$,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

$$X'_{k'}(I_T \otimes \Sigma_u^{-1})(vec(\Delta Y) - X vec(\hat{R}'_T)) = \frac{1}{2} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^{\gamma}} sign(\hat{R}'_T(i,j)) \tag{28}$$

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

$$\frac{1}{T_{k'}} X'_{k'}(I_T \otimes \Sigma_u^{-1})(vec(\Delta Y) - X vec(\hat{R}'_T)) = \frac{1}{T_{k'}} \frac{1}{2} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^{\gamma}} sign(\hat{R}'_T(i,j)) \tag{29}$$

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 $j \leq r$, $\frac{\lambda_{i,j,T}^{rank} T^{\frac{1}{2}(\gamma-1)}}{|\sqrt{T}\widetilde{R}(i,j)|^{\gamma}} \to \infty$. If $j > r$, $\frac{\lambda_{i,j,T}^{rank} T^{\gamma-1}}{|T\widetilde{R}(i,j)|^{\gamma}} \to \infty$

By KKT condition, if a zero element is estimated to be nonzero, then the equation (29) 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 3.3

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

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

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

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

We have shown the asymptotics of $\frac{1}{T}\Delta X C \Delta X'$ and $\frac{1}{T}UC\Delta X'$ in Lemma A.2. 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 3.2

Lemma A.3. *If error terms u_t in equation (1) are defined in Assumption 3.2, then the least squares estimate for Π is distributed as*

$$
\begin{aligned}
& vec\Big[\Big(Q(\widetilde{\Pi}-\Pi)Q^{-1}-[Q\Upsilon\Sigma_{z1}^{-1},0]\Big)D_T\Big] \\
= \quad & vec\Big[[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}Qw_tZ_{1,t-1}',\frac{1}{T}\sum_{t=1}^{T}Qw_tZ_{2,t-1}']\begin{bmatrix}\frac{1}{T}\sum_{t=1}^{T}Z_{1,t-1}Z_{1,t-1}' & \frac{1}{T^{3/2}}\sum_{t=1}^{T}Z_{1,t-1}Z_{2,t-1}' \\ \frac{1}{T^{3/2}}\sum_{t=1}^{T}Z_{2,t-1}Z_{1,t-1}' & \frac{1}{T^2}\sum_{t=1}^{T}Z_{2,t-1}Z_{2,t-1}'\end{bmatrix}^{-1}\Big] \\
\to_d \quad & \begin{bmatrix} N(0,\Sigma_{z1z1}^{-1}\otimes\Sigma_v) \\ vec\Big\{\Big((\Lambda\int_0^1 W_m dW_m'P')'+\sum_{j=1}^{\infty}\Gamma(j)\Big)\begin{bmatrix}0_{r\times(m-r)}\\ I_{m-r}\end{bmatrix} \\ \qquad\times\Big(\begin{bmatrix}0_{(m-r)\times r} & I_{m-r}\end{bmatrix}\Lambda(\int_0^1 W_mW_m'ds)\Lambda'\begin{bmatrix}0_{r\times(m-r)}\\ I_{m-r}\end{bmatrix}\Big)^{-1}\Big\}\end{bmatrix}
\end{aligned}
$$

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 $\Gamma(h)=\sum_{j=0}^{\infty}QD_{j+h}\Sigma_wD_j'Q'$.

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

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

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,

$$\frac{1}{T}\sum_{t=1}^{T}Qu_tZ_{1,t-1}' \;=\; \frac{1}{T}\sum_{t=1}^{T}Qw_tZ_{1,t-1}'+\frac{1}{T}\sum_{t=1}^{T}Q(\kappa(L)-\kappa(0))w_tZ_{1,t-1}'$$

with $\frac{1}{T}\sum_{t=1}^{T}Q(\kappa(L)-\kappa(0))w_tZ_{1,t-1}'\to_p\sum_{j=1}^{\infty}Q\kappa_j\Sigma_w\mathcal{X}_{j-1}'\equiv Q\Upsilon$. 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

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

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 3.2 can be referred to Lütkepohl (2007)

□

Proof of Proposition 3.1

The proof is similar to the proof of Theorem 3.2 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

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

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

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

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

Proof of Theorem 3.4

As in the proof of Theorem 3.2, we define such an objective function:

$$\begin{aligned}\Psi_T(E) &= \left\| vec(\Delta Y) - (\left[\begin{array}{cc} Y'_{-1}\hat{\beta}^{\dagger} & \Delta X^{p\prime} \end{array}\right] \otimes I_m) vec(\left[\begin{array}{cc} \alpha & B^p \end{array}\right] + \frac{1}{\sqrt{T}} E) \right\|^2_{I_T \otimes \Sigma_u^{-1}} \\ &+ \sum_{i=1}^{m}\sum_{j=1}^{r} \lambda_{i,j,T}^{rank} |\alpha(i,j) + \frac{1}{\sqrt{T}} E_0(i,j)| \\ &+ \sum_{k=1}^{p}\sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_{i,j,T}^{lag,k} |B_k(i,j) + \frac{1}{\sqrt{T}} E_k(i,j)|\end{aligned} \tag{30}$$

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

$$\begin{aligned}\Delta_T(E) &= \Psi_T(E) - \Psi_T(0) \\ &= vec(\frac{1}{\sqrt{T}}E)'(\left[\begin{array}{c} \hat{\beta}^{\dagger\prime} Y_{-1} \\ \Delta X^p \end{array}\right] \otimes I_m)(I_T \otimes \Sigma_u^{-1})(\left[\begin{array}{cc} Y'_{-1}\hat{\beta}^{\dagger} & \Delta X^{p\prime} \end{array}\right] \otimes I_m) vec(\frac{1}{\sqrt{T}}E) \\ &- 2vec(U)'(I_T \otimes \Sigma_u^{-1})(\left[\begin{array}{cc} Y'_{-1}\hat{\beta} & \Delta X^{p\prime} \end{array}\right] \otimes I_m) vec(\frac{1}{\sqrt{T}}E) \\ &+ \sum_{i=1}^{m}\sum_{j=1}^{r} \lambda_{i,j,T}^{rank} (|\alpha(i,j) + \frac{1}{\sqrt{T}} E_0(i,j)| - |\alpha(i,j)|) \\ &+ \sum_{k=1}^{p}\sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_{i,j,T}^{lag,k} (|B_k(i,j) + \frac{1}{\sqrt{T}} E_k(i,j)| - |B_k(i,j)|)\end{aligned} \tag{31}$$

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 3.2, we can similarly show that

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

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, $k = 1, \ldots, p$.

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 3.4 follows naturally.

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

□

Proof of Theorem 4.1

According to (5), the impulse responses matrices Φ_j are the elements of the upper left-hand $(m \times m)$ block of A^j. Therefore it is equivalent to show $\hat{A}^j$ are consistent estimates in the MA representation. With the notations defined in Section 2, the estimated coefficient matrices $\hat{A}$ are

$$
\hat{A} = \begin{bmatrix}
I_m + \hat{B}_1 + \hat{\Pi} & \hat{B}_2 - \hat{B}_1 & \cdots & \hat{B}_p - \hat{B}_{p-1} & -\hat{B}_p \\
I_m & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & I_m & 0
\end{bmatrix}
$$

In Lemma 3.1 we see that $\widetilde{S}_1$ is a consistent estimator of β', and according to Theorem 3.2, $\hat{R}'_T$ is a consistent rank estimator. Then

$$\begin{aligned}
\|\hat{\Pi} - \Pi\|_2 &= \|\hat{R}'_T\widetilde{S}'_1 - \alpha\beta'\|_2 \\
&= \|\hat{R}'_T\widetilde{S}'_1 - \alpha HH'\beta'\|_2 \\
&= \|\hat{R}'_T\widetilde{S}'_1 - \hat{R}'_T H'\beta' + \hat{R}'_T H'\beta' - \alpha HH'\beta'\|_2 \\
&= \|\hat{R}'_T(\widetilde{S}'_1 - H'\beta') + (\hat{R}'_T - R')H'\beta'\|_2 \\
&\leq \|\hat{R}'_T\|_2 \cdot \|\widetilde{S}'_1 - H'\beta'\|_2 + \|\hat{R}'_T - R'\|_2\|H'\beta'\|_2 = O_P(\frac{1}{\sqrt{T}})
\end{aligned}$$

according to Theorem 3.2 and Theorem 3.3 and $\|\hat{B}_k - B_k\|_2 = O_P(\frac{1}{\sqrt{T}})$ for $k = 1, \ldots, p$ by Theorem 3.3. Thus we have $\|\hat{A} - A\|_2 O_P(\frac{1}{\sqrt{T}})$.

In order to show $\|\hat{A}^j - A^j\|_2 \to_p 0$, we start with $p = 2$ as follows

$$\begin{aligned}
\|\hat{A}^2 - A^2\|_2 &= \|\hat{A}^2 - \hat{A}A + \hat{A}A - A^2\|_2 \\
&= \|\hat{A}(\hat{A} - A) + (\hat{A} - A)A\|_2 \\
&\leq \|\hat{A}\|_2 \cdot \|\hat{A} - A\|_2 + \|\hat{A} - A\|_2\|A\|_2 \\
&= O_p(\|\hat{A} - A\|_2) \to_p 0
\end{aligned}$$

where $\|A\|_2 = 1$. Now assume that $\|\hat{A}^{j-1} - A^{j-1}\|_2 = O_p(\|\hat{A} - A\|_2)$ for finite j, then we have,

$$\begin{aligned}
\|\hat{A}^j - A^j\|_2 &= \|\hat{A}^{j-1}\hat{A} - \hat{A}^{j-1}A + \hat{A}^{j-1}A - A^{j-1}A\|_2 \\
&= \|\hat{A}^{j-1}(\hat{A} - A) + (\hat{A}^{j-1} - A^{j-1})A\|_2 \\
&= O_p(\|\hat{A} - A\|_2)
\end{aligned}$$

Since $\hat{\Phi}_j$ are the elements of the upper left-hand $(m \times m)$ block of $\hat{A}^j$, it follows that $\|\hat{\Phi}_j - \Phi_j\|_2 \to_p 0$. □

Proof of Theorem 4.2

In Lemma 3.1 we see that $\hat{\Phi}_j$ is a consistent estimator of Φ_j, and consistent estimate of Σ_u is given by $\hat{\Sigma}_u$.With the notation from (21),

$$\begin{aligned}\|\hat{F}_h - F_h\|_2 &= \sum_{j=0}^{h-1} \|\hat{\Phi}_j \hat{\Sigma}_u \hat{\Phi}_j' - \Phi_j \Sigma_u \Phi_j'\|_2 \\ &= O_p(\|\hat{\Phi}_j - \Phi_j\|_2^2) + O_P(\|\hat{\Sigma}_u - \Sigma_u\|_2) \\ &= O_P(\frac{1}{T})\end{aligned}$$

The last equality follows from Theorem 4.1 and T-consistency of $\hat{\Sigma}_u$.

□

B Additional Results

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

Lemma 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 $\hat{\beta}^{\dagger\prime}$.*

$$T(\hat{\beta}_2\hat{\beta}_1^{-1} - \beta_0) \to_d (\int_0^1 W_{m-r}^* dW_r^*)'(\int_0^1 W_{m-r}^* W_{m-r}^{*'} ds)^{-1} \tag{32}$$

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 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 B.1 is more concentrated around 0. For general VECM, a similar result applies.

C Model estimation and specifications for simulations

C.1 Simulation result for model 1

Model 1: The experiments 7 and 8 in Chao and Phillips (1999) are a trivariate VAR with one lag and two cointegration vectors entering a single equation of the system. In their setting, the Monte Carlo study has demonstrated that their criterion performs well in small samples. Our model 1 is based on the same specification as the experiment 8, but consider different error structure. In addition to $\rho = 0.0$, we allow for strong cross-sectional dependence by choosing $\rho = 0.6$. Therefore we have the following specification which satisfies the standard assumptions,

$$\Delta Y_t = \alpha\beta' Y_{t-1} + B_1 \Delta Y_{t-1} + u_t \tag{33}$$

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

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

Table 8 reports the comparison of rank and lag selection results based on Model 1. The results indicate that lag selection performs well independent 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 ($T = 200$, $\rho = 0.0$)				**Model 1** ($T = 500$, $\rho = 0.0$)			
	$c = 1$	$c = 2$	$c = 3$		$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	100/95	100/100	96/100	$\gamma = 2.0$	100/99	100/100	100/100
$\gamma = 3.0$	98/100	80/100	59/100	$\gamma = 3.0$	100/100	100/100	99/100
$\gamma = 4.0$	80/100	50/100	24/100	$\gamma = 4.0$	100/100	87/100	62/100
$\gamma = 5.0$	57/100	22/99	10/98	$\gamma = 5.0$	88/100	50/100	20/100
Model 1 ($T = 200$, $\rho = 0.6$)				**Model 1** ($T = 500$, $\rho = 0.6$)			
	$c = 1$	$c = 2$	$c = 3$		$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	100/86	100/100	92/100	$\gamma = 2.0$	100/81	100/99	100/100
$\gamma = 3.0$	98/100	80/100	58/100	$\gamma = 3.0$	100/100	100/100	97/100
$\gamma = 4.0$	79/100	48/100	27/100	$\gamma = 4.0$	98/100	89/100	66/100
$\gamma = 5.0$	54/100	27/100	14/100	$\gamma = 5.0$	89/100	55/100	28/100

Table 8: Absolute numbers XX/YY of correct model selections by solving (16) and (17) 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.

C.2 Model specifications

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 \\
-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 \\
-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 \\
-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 \\
-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 \\
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 \\
-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 \\
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 \\
-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 \\
-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 \\
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 \\
-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 \\
-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 \\
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 \\
-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 \\
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 \\
1.520876 & 0.133889 & -0.8365487 & 2.135475 & -2.056529 & 0.9585998 & 0.6852929 & -0.5481826
\end{bmatrix}
$$

C.3 Comparison of different estimation methods

$T = 200$	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 9: Comparison of different estimation methods for Model 1

$T = 200$	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 10: Comparison of different estimation methods for Model 2

$T = 200$	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 11: 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 12: Comparison of different estimation methods for Model 4

D Additional Empirical Results

$$\hat{\alpha} = \begin{bmatrix} -1.753 & 0.047 & 0 & \dots & 0 \\ -0.170 & -1.405 & 0 & \dots & 0 \\ -0.185 & -0.339 & & & \\ -0.445 & -0.825 & & & \\ -0.301 & -1.002 & & & \\ -0.219 & -0.819 & & & \\ -0.256 & -0.909 & & & \\ -0.232 & -0.822 & \vdots & \ddots & \vdots \\ -0.382 & -0.719 & & & \\ -0.500 & -0.791 & & & \\ -0.253 & 0.044 & & & \\ -0.370 & -0.991 & & & \\ -0.386 & -0.999 & & & \\ -0.224 & -0.736 & & & \\ -0.422 & -0.975 & & & \\ -0.421 & -0.966 & & & \\ -0.380 & -0.981 & 0 & \dots & 0 \end{bmatrix}$$

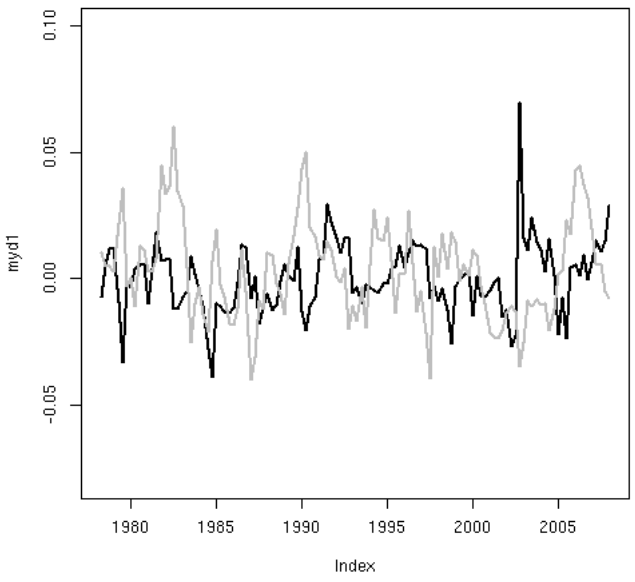

Figure 5: The time-varying pattern of two cointegration factors, with η_1 in black and η_2 in gray.

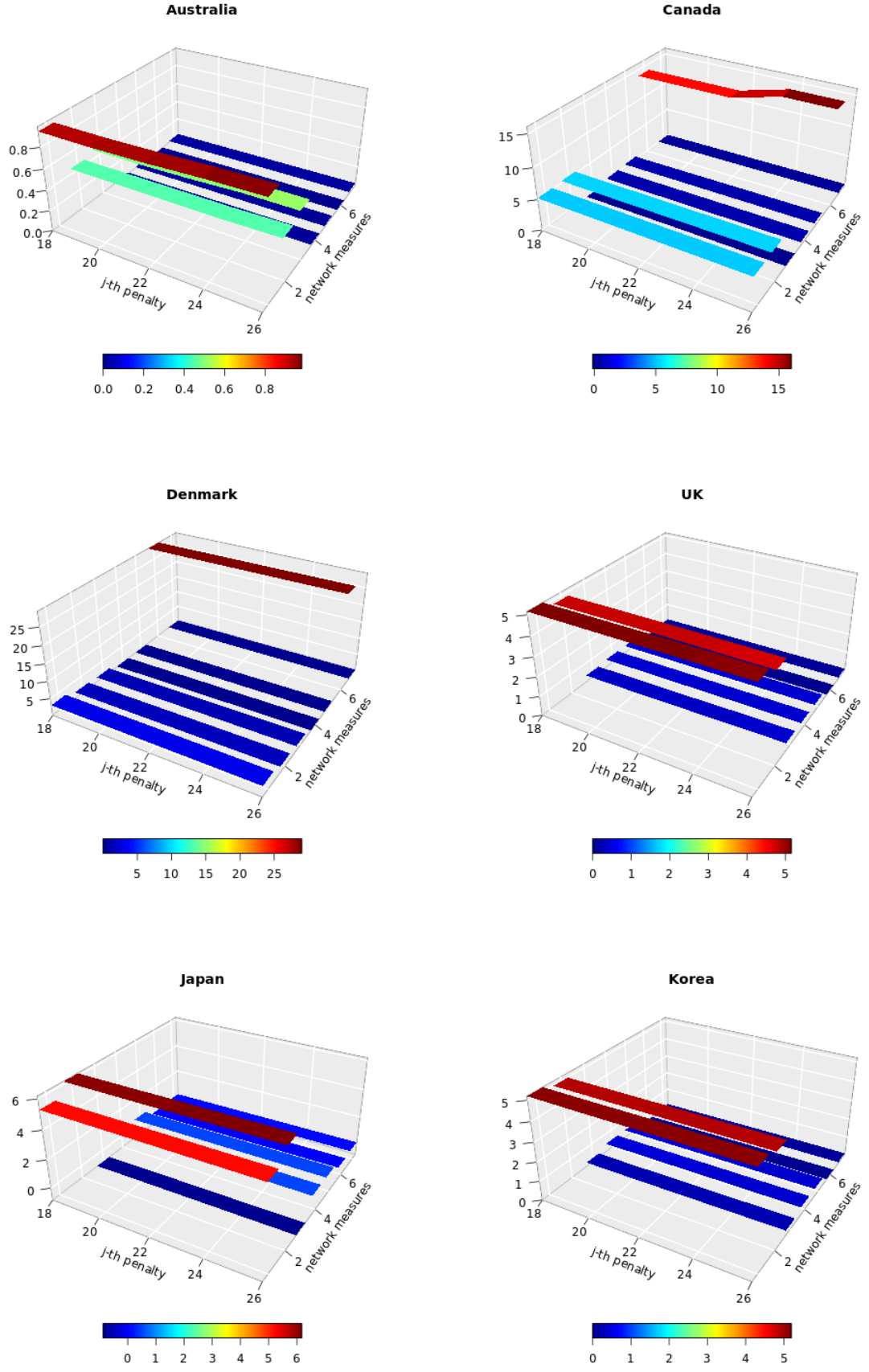

Figure 6: This figure graphically shows seven network measures numbered as follows: 1.$C_{from,i}$, 2.$C_{to,i}$, 3.$C_{net,i}$, 4.$indeg(i)$, 5.$outdeg(i)$, 6.$Bet(i)$, 7.$Clos(i)$ for a range of tuning parameters.

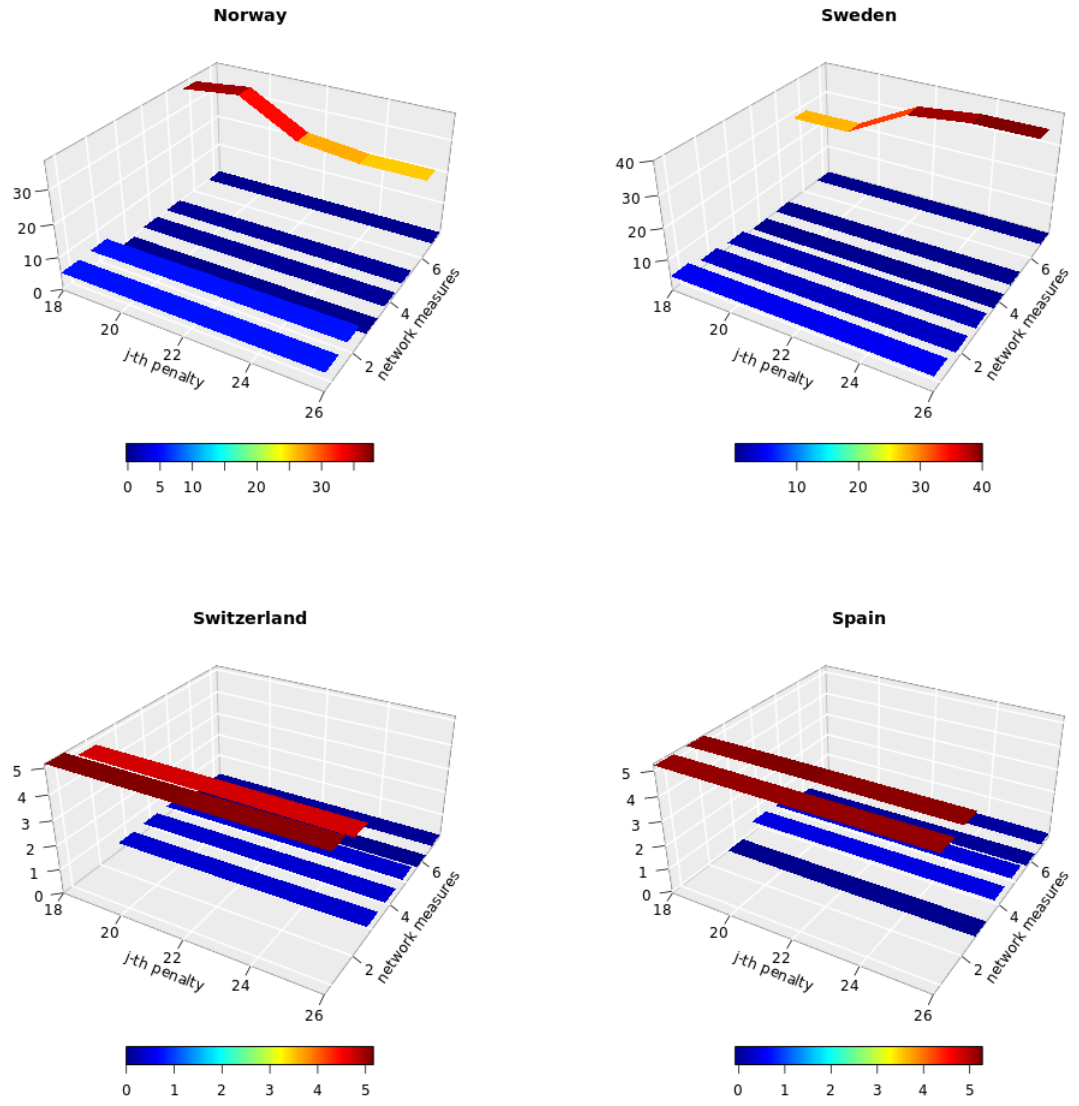

Figure 7: This figure graphically shows seven network measures numbered as follows: 1.$C_{from,i}$, 2.$C_{to,i}$, 3.$C_{net,i}$, 4.$indeg(i)$, 5.$outdeg(i)$, 6.$Bet(i)$, 7.$Clos(i)$ for a range of tuning parameters.

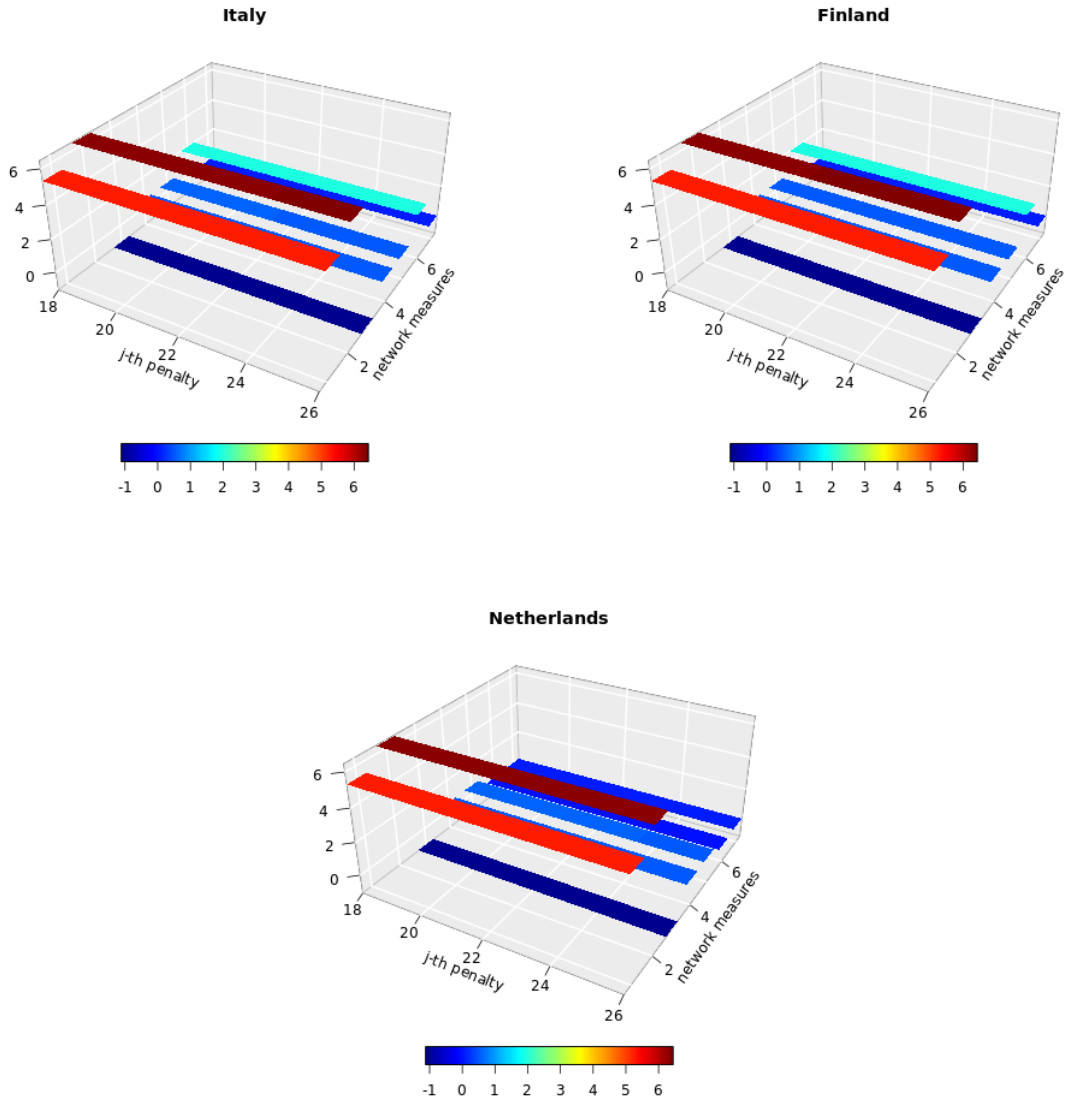

Figure 8: This figure graphically shows seven network measures numbered as follows: 1.$C_{from,i}$, 2.$C_{to,i}$, 3.$C_{net,i}$, 4.$indeg(i)$, 5.$outdeg(i)$, 6.$Bet(i)$, 7.$Clos(i)$ for a range of tuning parameters.