Tailored Lasso for Moderate Dimensional VECM and Exchange Rates

Shi Chen* Melanie Schienle†

February 18, 2019

Abstract

In this paper, we propose a tailored Lasso-type technique for consistent and numerically efficient model selection which is feasible for both, standard low but also higher dimensions. The proposed adaptive shrinkage method allows for model choice and direct estimation in one step. We derive the corresponding asymptotic results for model selection and also propose a refinement strategy to obtain final estimates. In a comprehensive simulation study we shown convincing finite sample performance of our techniques. For a system of FX rates, we highlight the advantages of the tailored elementwise adaptive Lasso procedure.

JEL classification: C32, C52, C53, F31
Keywords: cointegration, VECM, exchange rate, Lasso, model selection

*Karlsruher Institut für Technologie, Lehrstuhl für Ökonometrie und Statistik, Blücherstr.17, 76185 Karlsruhe, Germany. Email: shi.chen@kit.edu.

†Karlsruher Institut für Technologie, Lehrstuhl für Ökonometrie und Statistik, Blücherstr.17, 76185 Karlsruhe, Germany. Email: melanie.schienle@kit.edu.

1 Introduction

Many financial and natural systems are dynamic, multi-dimensional and often contain a large number of non-stationary potentially cointegrated components. Generally, the standard tool to handle such multivariate time-series has been the vector error correction model (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., 2012) 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., 2015) 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 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 asymptotics 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 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. Our FX application illustrates that such refinements can make a difference in practice.

In this paper, 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 are actually prevalent for the system. In the same way, a small number of VAR lags are considered sufficient for a parsimonious model specification, but within this maximum lag range, our model selection technique is independent from the lag ordering. In this sense, the problem is assumed to be sparse. We show consistency of the variable selection by the proposed Lasso-VECM estimator and derive its asymptotic properties for inference. For refined estimation in particular in larger dimensional finite samples, we provide a refined estimation strategy and derive its statistical properties. Moreover, with only linear computational complexity, all procedures remain computationally tractable also for higher dimensions. 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 si-

mulation study shows the effectiveness of the proposed techniques in finite samples. This is also illustrated by the empirical study for FX data where we find a superior performance of predictions based on our tailored elementwise adaptive lasso techniques.

Our work builds on the vast literature of VECM as summarized e.g. in Lütkepohl (2007) as well as on results for Lasso techniques in the standard *i.i.d.* case. Lasso was proposed by Tibshirani (1996) and its asymptotic properties were first studied in Knight and Fu (2000). The adaptive Lasso by Zou (2006) improved on the selection properties by penalizing different variables differently. Yuan and Lin (2006) introduce group-Lasso which allows for simultaneous exclusion and inclusion of certain variables. For the Lasso optimization, there are several standard solution algorithms such as the coordinate descent (Friedman et al., 2007; Friedman et al., 2010, or others), the interior point method (Koh et al., 2007), or the orthant-wise limited-memory Quasi-Newton optimizer (Andrew and Gao, 2007). Our proposed technique is also related to recent literature which uses Lasso for model determination in a stationary high-dimensional VAR context (Kock and Callot, 2015) but cannot handle nonstationary components. For non-stationary time series, there also exist some empirical and simulation work employing penalizing algorithms for VECM without mathematical proofs, see e.g. Signoretto and Suykens (2012), Wilms and Croux (2016). A similar setting to ours has been investigated in a recent theoretical paper by Liao and Phillips (2015). Our proposed two-step approach, however, differs in linearity of the objective Lasso function and corresponding numerically efficient solution algorithm which both lead to feasibility advantages in particular in higher dimensions. 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. For applications, this can be key to augmented forecasting results as illustrated in the studied FX case.

The paper is organized as follows. Section 2 presents the model setup. Section 3 gives the main results on the limit theory of the model selection consistency. Section 4 presents extensions of the proposed approach. In particular, we show strategies for refined estimation and derive results when the error terms are weakly dependent. Section 6 presents the empirical findings for FX rates. All proofs are contained in the Appendix. 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.

2 Model and Estimation

In order to illustrate the main ideas of the proposed Lasso methodology, we first derive our Lasso objective function in a simple setting for a known fixed VAR with one lag. Thus model determination here only consists of cointegrating rank selection and estimation. We denote this setup as special case described in Subsection 2.1. Results are of independent interest, as such models are widely used in the applied literature. In Subsection 2.2, we generalize the setting to a general unknown VAR with unknown general lag order which then also enters the model selection problem. Thus complete model specification then amounts to both rank and lag order determination.

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

2.1 Special case

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

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

with equivalent VECM representation

$$\Delta Y_t \;\; = \;\; \Pi Y_{t-1} + u_t \tag{2}$$

for $t = 1, \ldots, T$, where $\Pi = A_1 - I_m$ 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 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, so only the space of cointegration relations is identified up to rotation but not β.

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

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

In Section 4.2, we show how Assumption 2.1 can be generalized admitting linear forms of weak dependence. Such a general setting requires changes in the Lasso procedure and leads to different statistical properties of the modified technique.

In this setting, VECM determination reduces to selection of the correct cointegration rank. Our shrinkage selection procedure is based on an available consistent pre-estimate of the cointegration matrix Π. It is well known, that for the model setting in (2) and Assumption 2.1 the standard least squares estimator

$$\widetilde{\Pi} = \Big(\sum_{t=1}^{T} \Delta Y_t Y_{t-1}'\Big)\Big(\sum_{t=1}^{T} Y_{t-1} Y_{t-1}'\Big)^{-1} \tag{3}$$

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

The distribution of $\widetilde{\Pi}$ relies on a Q-transformation of Y_t, which allows to disentangle stationary and nonstationary components. It pre-multiplies all elements in (2) from the left with the specific matrix Q defined as follows

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

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

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

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

[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 $M'M_\perp = 0$.

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

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

$$\begin{aligned} \widetilde{\Pi} &= \widetilde{R}'\widetilde{S}' \\ &= \begin{bmatrix} \widetilde{R}'_{1,m\times r} & \widetilde{R}'_{2,m\times(m-r)} \end{bmatrix} \begin{bmatrix} \widetilde{S}'_{1,r\times m} \\ \widetilde{S}'_{2,(m-r)\times m} \end{bmatrix} \end{aligned} \tag{4}$$

where $\widetilde{S}$ is an orthonormal matrix, i.e. $\widetilde{S}'\widetilde{S} = I$. $\widetilde{R}$ is an upper triangular matrix [3]and further properties of this decomposition can be found in Stewart (1984). Column-pivoting orders columns in R according to size putting zero-columns at the end.[4] Since $\widetilde{\Pi}$ is a matrix of full-rank and also a consistent estimate of Π, the lower diagonal elements of the last $(m-r)$ columns of the matrix $\widetilde{R}'$ are expected to be small, converging to zero asymptotically at unit root speed T. This is shown in Lemma 2.1, where we derive convergence results of the QR-decomposition components $\widetilde{R}$ and $\widetilde{S}$ from the least squares pre-estimate $\widetilde{\Pi}'$.

Lemma 2.1. *Let Assumption 2.1 hold for $\widetilde{\Pi}$ in* (3)*. We denote by $\widetilde{R}'_1$ the first r and by $\widetilde{R}'_2$ the last $m-r$ columns of $\widetilde{R}'$ in the QR-decomposition* (4) *of $\widetilde{\Pi}'$. Let β be orthonormal and H be some $(r\times r)$-orthonormal matrix. Then*

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

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

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

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

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

Lemma 2.1 clearly shows that the last $m - r$ columns of $\widetilde{R}'$ converge to zero at rate T, faster than the $\sqrt{T}$- rate of the first r stationary columns. We exploit this idea in constructing adaptive weights for a model selection consistent Lasso procedure, which put a faster diverging penalty on true zero singular values of Π and less on the non-zero ones corresponding to the underlying stationary components. In particular, we expect that zero columns in R' can be easily detected by adaptive Lasso, as non-zero columns estimated as close to zero would converge slower than true zero components approach zero according to Lemma 2.1. Therefore relating penalties in adaptive Lasso to inverses of these initial estimates shrinks true zero components faster to zero than the other ones, which results in a higher penalty for the true zero parts and the detection of the appropriate basis for the cointegration space. Then this adaptive penalty causes the number of non-zero columns in the penalized estimate of R' to produce a consistent estimate for the rank r of Π. Hence elements $\widehat{R}(i,j)$ of $\widehat{R}$ minimize the following criterion over all $R(i,j)$ for $i, j = 1, \ldots, m$

$$
\sum_{t=1}^{T} \parallel \Delta Y_t - R' \widetilde{S}' Y_{t-1} \parallel_{\Sigma_u^{-1}}^2 + \sum_{i,j=1}^{m} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^\gamma} |R(i,j)| \tag{5}
$$

where $\widetilde{R}(i,j)$ is the (i,j)th element of an un-penalized pre-estimate $\widetilde{R}$ generated from the QR-decomposition of $\widetilde{\Pi}'$ in (4). The penalization parameter λ and the weight γ for adaptiveness are fixed and in practice pre-determined in a data-driven way. See the simulation and application in Sections 5 and 6 for details. We then obtain an estimate of the true cointegration rank $\hat{r}$ from (5) as $\hat{r} = \text{rank}(\hat{R})$, where $\text{rank}(\hat{R})$ equals the number of non-zero columns in $\hat{R}'$. Another advantage of such an objective function is that even non-zero columns in $\hat{R}'$ can still have zero elements, which exploits the sparsity structure of R sufficiently and thus leads to extra efficient gains. We use a GLS-type loss function in (5) and also in the subsequent subsection for efficiency purposes in general cases of Σ_u. All are operationalized with the corresponding FGLS approach by minimizing the corresponding empirical norm with pre-estimated $\widetilde{\Sigma}_u^{-1}$.

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

Moreover, compared with existing literature for our setting, our approach features several advantages: Firstly, the employed QR-decomposition is real-valued without further constraints on the matrix $\widetilde{\Pi}$. Thus the Lasso criterion (5) only contains real-valued elements and can be minimized with standard optimization techniques. In higher-dimensions, however, a corresponding eigenvalue decomposition would most likely contain complex values leading to a non-standard harmonic function optimization problem in a respective Lasso objective function. Secondly, the form of the Lasso objective functions is linear in coefficients and therefore straightforward to implement relying on available numerically efficient standard Lasso procedures. So our method is direct and ready to use. And thirdly, the form of our objective function directly allows to employ sparsity constrains for efficient estimation which seems important in particular in higher dimensional settings.

2.2 General case

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

$$\Delta Y_t \;=\; \Pi Y_{t-1} + B_1 \Delta Y_{t-1} + \cdots + B_P \Delta Y_{t-P} + u_t \tag{6}$$

for $t = 1, \ldots, T$. As before, the dimension of $\{Y_t\}$ is m, the rank of Π is $r < m$. We set the maximum possible lag length P as sufficiently large but fixed independent of T, such that it is an upper bound for the true lag p, i.e. $p < P$. In this case, $B_{p+1}, \ldots, B_P$ are all zero matrices. For sparsity, additionally $P/p = c_2$ with $c_2 \gg 1$.

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

Assumption 2.2. *1. The roots for* $|(1-z)I_m - \Pi z - \sum_{j=1}^{p} B_j(1-z)z^j| = 0$ *is either* $|z| = 1$ *or* $|z| > 1$.

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

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

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

For estimation purposes, we rewrite the general VECM defined in (6), for $t = 1, \ldots, T$, in matrix notation as

$$\Delta Y = \Pi Y_{-1} + B\Delta X + U \tag{7}$$

where $\Delta Y = [\Delta Y_1, \ldots, \Delta Y_T]$, $Y_{-1} = [Y_0, \ldots, Y_{T-1}]$, $B = [B_1, \ldots, B_P]$, $\Delta X = [\Delta X_0, \ldots, \Delta X_{T-1}]$ where $\Delta X_{t-1} = \left[\Delta Y'_{t-1}, \ldots, \Delta Y'_{t-P}\right]'$ 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 2.2, it holds by Lemma 1 in Toda and Phillips (1993)

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{[Ts]} \Gamma_t \to_p B_\Gamma(s) \tag{8}$$

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

$$\Sigma_{\Gamma\Gamma} = \begin{pmatrix} \Sigma_{z1z1} & \Sigma_{z1\Delta x} \\ \Sigma_{\Delta xz1} & \Sigma_{\Delta x\Delta x} \end{pmatrix} \tag{9}$$

The least squares estimate for (7) is denoted by $[\widetilde{\Pi}_{ls}, \widetilde{B}_{ls}]$, which will be used to get the consistent estimate $\widetilde{\Sigma}_u = \frac{1}{T-mP+1}(\Delta Y - \widetilde{\Pi}_{ls}Y_{-1} - \widetilde{B}_{ls}\Delta X)(\Delta Y - \widetilde{\Pi}_{ls}Y_{-1} - \widetilde{B}_{ls}\Delta X)'$ of Σ_u (see e.g. Lütkepohl, 2007).

For model selection, we disentangle the joint lag-rank selection problem by employing a Frisch-Waugh-idea in the VECM model (7). With this, we obtain two independent criteria for lag and rank choice which can be computed separately. For rank selection, the partial least squares pre-estimate $\widetilde{\Pi}$ can be obtained from the corresponding partial model when removing the effect of ΔX in ΔY and Y_{-1} by regressing $\Delta Y M$ on $Y_{-1}M$ with $M = I_T - \Delta X'(\Delta X \Delta X')^{-1}\Delta X$. Thus it is

$$\widetilde{\Pi} = \left(\Delta Y M Y'_{-1}\right)\left(Y_{-1} M Y'_{-1}\right)^{-1} \tag{10}$$

Lemma 2.2. *Under Assumptions 2.1 and 2.2, the partial least squares estimate* $\widetilde{\Pi}$ *defined*

in (10) satisfies

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

where $D_T = diag(\sqrt{T}I_r, TI_{m-r})$, $\Sigma_v = Q\Sigma_u Q'$, $Z_{-1} = \beta' Y_{-1}$, $\frac{1}{T}Z_{-1}MZ_{-1}' \rightarrow_p \Sigma_{z1z1.\Delta x} = \Sigma_{z1z1} - \Sigma_{z1\Delta x}\Sigma_{\Delta x\Delta x}^{-1}\Sigma_{\Delta xz1}$ *with all the compoment covariance matrices defined in* (9)*;* $W_{m-r}^{\dagger} = (\alpha_\perp'\Sigma_u\alpha_\perp)^{-\frac{1}{2}}[0_{(m-r)\times r}, I_{m-r}]\Sigma_v^{\frac{1}{2}}W_m$*, 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* (18) *and* (19) *in the proof.*

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

Lemma 2.2 shows that $\widetilde{\Pi}$ is a consistent estimate. Thus we can employ the idea of the previous subsection for rank selection and separate the problem into stationary and nonstationary parts as in the special case. We thus obtain for the components of the QR-decomposition $\widetilde{\Pi} = \widetilde{R}'\widetilde{S}'$:

Lemma 2.3. *Let Assumptions 2.1 and 2.2 hold for* $\widetilde{\Pi}$ *in* (10)*. 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* (4) *of* $\widetilde{\Pi}'$ *defined in* (10)*. Let* β *be orthonormal and* H *be a* $(r\times r)$*-orthonormal matrix.*

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

where $\frac{1}{T}\beta' Y_{-1}MY_{-1}'\beta \rightarrow_p \Sigma_{z1z1.\Delta x}$ *and* $\Sigma_{z1z1.\Delta x}$ *is defined as in Lemma 2.2.*

Thus from Lemma 2.2 and 2.3, we can construct a corresponding adaptive Lasso procedure as an analogue to (5) in vector form. Hence components $\hat{R}(i,j)$ of $\hat{R}$ minimize the following criterion over all $R(i,j)$ for $i,j = 1,\ldots,m$

$$\|vec(\Delta YM) - (MY_{-1}'\widetilde{S}\otimes I_m)vec(R')\|_{I_T\otimes\Sigma_u^{-1}}^2 + \sum_{i,j=1}^{m}\frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^\gamma}|R(i,j)| \qquad (11)$$

where now $\widetilde{R}(i,j)$ is from the QR-decomposition of $\widetilde{\Pi}'$ in the partial model (10). We choose the cointegration rank as $\hat{r} = \text{rank}(\hat{R})$, where $\text{rank}(\hat{R})$ is the number of non-zero columns in $\hat{R}'$.

Likewise, for independent lag selection, the effect of the nonstationary term Y_{-1} in (7) must be filtered out in ΔY and ΔX for unbiased estimation in the partial model via regression of ΔYC on ΔXC with $C = I_T - Y'_{-1}(Y_{-1}Y'_{-1})^{-1}Y_{-1}$. Thus we obtain $\hat{B}$ as minimizing the following objective function over all components $B_k(i,j)$ for $k = 1, \ldots, P$ and $i, j = 1, \ldots, m$

$$||vec(\Delta YC) - (C\Delta X' \otimes I_m)vec(B)||^2_{I_T \otimes \Sigma_u^{-1}} + \sum_{k=1}^{P}\sum_{i,j=1}^{m} \frac{\lambda_{i,j,T}^{lag,k}}{|\check{B}_k(i,j)|^\gamma}|B_k(i,j)| \tag{12}$$

for fixed tuning parameters $\lambda_{i,j,T}^{lag,k}, \gamma$, where γ here and in the rank selection (11) might differ. Moreover, the pre-estimate $\check{B}$ in the adaptive Lasso weight can be taken from the partial least squares estimate $\widetilde{B} = (\Delta YC\Delta X')(\Delta XC\Delta X')^{-1}$ due to consistency. Though in practice, especially with larger dimensions and lags, multicollinearity effects in ΔX are quite likely to occur which cause the least squares estimate to become numerically instable. Therefore we also consider a robust ridge type pre-estimate $\widetilde{B}^R$ as $\check{B}$, which can be obtained from

$$\begin{aligned} \widetilde{B}^R \quad &= \arg\min \|vec(\Delta YC) - (C\Delta X' \otimes I_m)vec(B)\|^2 \\ &\quad + \nu_T \textstyle\sum_{k=1}^{P}\sum_{i,j=1}^{m} |B_k(i,j)|^2 \end{aligned} \tag{13}$$

The following Theorem 2.1 shows that this pre-estimate is also consistent for appropriate choices of tuning parameters

Theorem 2.1. *If the tuning parameter ν_T in the ridge regression (13) satisfies $\frac{\nu_T}{\sqrt{T}} \to_p 0$, then $\sqrt{T}(\widetilde{B}^R - B) = O_p(1)$ under Assumptions 2.1 and 2.2.*

As in the case of rank selection, a lag k should be included into the model, whenever $\hat{B}_k$ from the Lasso selection (12) is different from zero. Thus, in contrast to other model selection criteria, a Lasso-type procedure allows for the inclusion of non-consecutive lags, which we consider an additional advantage of the procedure. We obtain an estimate $\hat{p}$ of the true lag length from (12) as $\hat{p} = \max_{1 \le k \le P}\{k|\hat{B}_k \neq 0\}$.

Note that the residual transformation C in the lag selection criterion (12) 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 multicollinearity 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 below)

3 Technical Results

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

3.1 Model Selection Consistency for special VECM

The following theorem derives the statistical properties of the estimate from our adaptive Lasso procedure (5) for special VECM.

Theorem 3.1. *Suppose that* $\lambda_{i,j,T}^{rank}/\sqrt{T} \to 0$ *and* $T^{\frac{1}{2}(\gamma-1)}\lambda_{i,j,T}^{rank} \to \infty$ *and* $\Pi' = SR$ *is a QR decomposition with column pivot. Then under Assumptions 2.1 the objective function (5) yields:*

1. $\lim_{T\to\infty} \mathbb{P}(\mathcal{A}_T^* = \mathcal{A}) = 1$
   where $\mathcal{A}$ *is the set of indices for the non-zero elements of* $vec(R')$, $\mathcal{A}_T^*$ *is the set of indices for the non-zero elements of* $vec(\hat{R}_T')$ *derived in* (5).

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

Theorem 3.1 shows that our method is consistent in variable selection i.e., it chooses the right rank and the correct sparse pattern with probability one. This is our primary and main concern. Note that the weight function in the adaptive Lasso procedure is crucial to achieve this property.

Additionally, the second part of the theorem gives the asymptotic distribution of the adaptive Lasso-VECM estimate. It is asymptotically unbiased converging to a normal distribution at the standard stationary speed $\sqrt{T}$. The complicated structure of the variance matrix is due to the sparse structure of R in our Lasso procedure. However, this estimate suffers from endogeneity bias caused by the naive estimate of β. More specifically, the bias in estimating β from the least squares estimate $\widetilde{\Pi}$ depends on the

term $\int_0^1 dW_m(s)W_m(s)'$, in which the integrand $W_m(s)$ and the differential part $dW_m(s)$ are the same Brownian motion, thus dependent. The bias could be further decreased if had the form $\int_0^1 dW_1(s)W_2(s)'$ with W_1, W_2 independent Brownian motions. The latter can be achieved by reduced rank regression, see Anderson (2002) for detailed asymptotics. Therefore, it is recommended to update β after obtaining a consistent estimate for r. For details we refer to Subsection 4.1.

If we treat each column of R' as a group and apply adaptive group Lasso, with a similar proof we can show that the right rank can still be estimated consistently. This has been studied in a general set-up in Liang and Schienle (2019). Though, in this case essential parts of the sparse structure are neglected which can led to inferior finite-sample estimation and prediction results in moderate dimensions. We highlight the advantage of the proposed elementwise procedure in detail in the empirical section 6.

3.2 Model Selection Consistency for general VECM

First, we show the result for the cointegrating rank selection according to criterion (11) which uses the residual transformation M in order to focus on the respective partial effect in the general VECM. The structure of the result resembles the one of the special case.

Theorem 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 2.2 with the same notation for* $\mathcal{A}$ *as in Theorem 3.1 the objective function (11) yields*

1. $\lim_{T\to\infty} \mathbb{P}(\mathcal{A}_T^* = \mathcal{A}) = 1$
   where $\mathcal{A}_T^*$ *is index set of the non-zero elements of* $vec(\hat{R}')$ *in* (11).

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 4.1 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 (12) with adaptive weights from a ridge regression pre-estimate $\widetilde{B}^R$. In this way, we account for prevalent multicollinearity effects in particular in settings with higher dimensions and large lag lengths.

Theorem 3.3. *Suppose that* $\lambda_{i,j,T}^{lag,k}/\sqrt{T} \to 0$ *and* $T^{\frac{1}{2}(\gamma-1)}\lambda_{i,j,T}^{lag,k} \to \infty$. *Then the lag objective function (12) yields:*

1. $\lim_{T\to\infty} \mathbb{P}(\mathcal{B}_T^* = \mathcal{B}) = 1$;
   where $\mathcal{B}$ *is the set of indices for the non-zero elements of* $vec(B)$, $\mathcal{B}_T^*$ *is the set of indices for the non-zero elements of* $vec(\hat{B})$ *in* (12)

2. $\sqrt{T}vec(\hat{B}_T' - B')_{\mathcal{B}} \to_d \mathrm{N}(0, (\Sigma_{\Delta x\Delta x.z1} \otimes \Sigma_u^{-1})_{\mathcal{B}}^{-1}(\Sigma_{\Delta x\Delta x.z1} \otimes \Sigma_u^{-1})_{\mathcal{B}}(\Sigma_{\Delta x\Delta x.z1} \otimes \Sigma_u^{-1})_{\mathcal{B}}^{-1})$
   where $\Sigma_{\Delta x\Delta x.z1} = \Sigma_{\Delta x\Delta x} - \Sigma_{\Delta xz1}\Sigma_{z1z1}^{-1}\Sigma_{z1\Delta x}$ *with all the compoment covariance matrices defined in* (9).

Thus lag selection is consistent i.e., the true lags are selected with probability 1 even if they are non-consecutive. For estimation of the coefficients in the relevant lag components, as in the case for the rank, we find asymptotic normality and unbiasedness at the standard stationary speed. Different to the rank selection result in Theorem 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.

4 Extensions

4.1 Refined model estimation in higher dimensions

With our proposed adaptive Lasso techniques, we can select the true model with probability one for sufficiently many observations. Although both model selection criteria (11) and (12) also yield consistent estimates for the coefficients of appropriate variables, there is, however, substantial room for improvement on the estimation side in particular in finite samples for higher dimensions. For pure model estimation in higher dimensions, we therefore suggest a refined procedure for α and B_k with $k \in \{1, \ldots, p\}$ which is still of Lasso type but no longer adaptive. With a focus on model estimation, given the pre-selected rank and lag, we propose a pure Lasso procedure rather than an adaptive variant. While the latter is targeted at consistent model selection, a pure Lasso estimate performs better in estimation and prediction (see Bühlmann and Van De Geer (2011) for the comparison of different variants of Lasso).

Besides, we use an improved estimate $\tilde{\beta}^{\dagger}$ of β from reduced rank regression (see Ahn and Reinsel (1990) and Anderson (2002)), which does not suffer from endogeneity bias and

yields improved finite sample performance. Please note, that generally $\tilde{\beta}^\dagger$ an efficient estimate of $\tilde{\beta}^\dagger$ relies on a precise estimate for the rank by matrix perturbation theory, as well as a consistent estimate for the lag p. Therefore in particular in higher-dimensional sparse settings, it can only be employed in the estimation refinement step and is no option for the pre-step in model selection.

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

$$\sum_{t=1}^{T} ||\Delta Y_t - \alpha \tilde{\beta}^{\dagger\prime} Y_{t-1} - \sum_{k=1}^{p} B_k \Delta Y_{t-k}||^2_{\Sigma_u^{-1}} + \sum_{i=1}^{m}\sum_{j=1}^{r} \lambda^{rank}_{i,j,T} |\alpha(i,j)| + \sum_{k=1}^{p}\sum_{i,j=1}^{m} \lambda^{lag,k}_{i,j,T} |B_k(i,j)| \tag{14}$$

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 (14).

We show that with appropriate choices of tuning parameters, the penalized estimates from (14) are consistent and yield the same asymptotic variance as the ones from reduced rank regression, while its solution is sparse in finite samples and thus improves the mean squared error in general. Though as the simulations in Section 5 will confirm, their finite-sample performance, however, is superior in particular for estimation but also for prediction.

Theorem 4.1. *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 (14) under Assumptions 2.1 and 2.2 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 4.1 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 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.

4.2 Model Selection with Dependent Error Terms

In this section 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 4.1. *In the special VECM as* (2) *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.*

Assumption 4.1 is stronger than absolute summability due to the convergence of unit root processes.

Lemma 4.1. *Under Assumption 4.1, the least squares estimate for* Π *in* (2) *is biased and satisfies*

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

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

The term Υ measures the correlation between u_t and $Z_{1,t-1}$ due to the auto-correlation of u_t under Assumption 4.1.

Define $\Xi = \begin{bmatrix} \beta' \\ \beta'_\perp \end{bmatrix}$ as in the proof for Lemma 2.1, we have

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

By a similar argument as for Lemma 2.1, we can conclude that

Lemma 4.2. *By the same notation as in Lemma 2.1 and under Assumption 4.1, the following results hold:*

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

Due to the bias term, we can't expect that the selection result from (5) is consistent element-wise, but consistency in rank could still hold when the penalty term is modified. The estimate $\hat{R}$ is obtained by minimizing the follwing objective function row-wise in $R(i,)$ for $i = 1, \ldots, m$

$$\sum_{t=1}^{T} \parallel \Delta Y_t - R'\widetilde{S}'Y_{t-1} \parallel_2^2 + \sum_{i=1}^{m} \frac{\lambda_{i,T}^{rank}}{||\widetilde{R}(i,)||_2^{\gamma}} ||R(i,)||_2 \tag{15}$$

Different from (5), we penalize each row in R as a group, similar to Yuan and Lin (2006), Wang and Leng (2008). Therefore, there could be zero and non-zero rows in $\hat{R}$, but non-zero rows have no zero elements. By Lemma 4.2, the penalty on the first r rows of R would be bounded and the penalty on the last $m - r$ rows explodes. Thus consistency of the estimate from (15) in rank selection is expected. Besides, the first term in (15) is equivalent to the ordinary least squares problem rather than a generalized least squares because we penalize the each row in R as a whole. The statistical property is given in Proposition 4.1.

Proposition 4.1. *Given Assumption 4.1, 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* (15) *is consistent in selecting the right rank.*

When the dimension is higher, the variance of $\hat{R}$ from (5) generally increases due to the non-sparse structure within non-zero rows of $\hat{R}$.

5 Simulations

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

But in particular, we focus on cases up to dimension eight and sixteen with a thorough simulation study of model selection quality as well as the estimation and forecast fit. Such higher dimensional specifications are not feasible with available standard techniques and provide a substantial generalization to the common bivariate illustrations in this literature.

In all model specifications we consider independent multivariate Gaussian innovations with covariance matrix $\Sigma_u = [\rho^{|i-j|}]_{i,j=1}^m$ for two particular cases $\rho = 0.0$ and $\rho = 0.6$. Thus our specifications include cases of strong cross-sectional dependence. The chosen vanishing pattern of correlations corresponds e.g. to increasing geographical distance in the case of the FX application in Section 6. For these settings, we use the general FGLS-type empirical versions of the objective functions (11) and (12) for model selection with least squares estimate $\widetilde{\Sigma}_u$ for Σ_u.

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

For transparency, we report all results dependent on the choice of tuning parameters γ and λ in the adaptive Lasso procedure. Thus for each setting, we show all results on a two-dimensional grid of $\lambda = cT^{1/2-\varepsilon}$ and γ where $\varepsilon = 0.1$ and c takes all integers from 1 to 3 and γ ranges from 2 to 5 in steps of 1. We use the same penalty λ for both rank and lag selection, which could in practice be refined with different tuning parameters for each criterion. Although lag and rank selection work independently, we found that 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 (11). In the literature, BIC is a standard way to choose tuning parameters. For comparison, we mark the BIC-selection of (γ, c) in the Tables by underlining respective values. They are obtained as minimizing the following criteria:

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

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

Simulations for model selection are done in `R`. Lasso is implemented with the package

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

For the standard three dimensional case, we choose a setting considered in Chao and Phillips (1999) for comparison purposes. For the higher dimensions, at each level of model complexity with given dimension, cointegration rank and lag length, our simulation settings are randomly chosen from all possible VECM specifications satisfying the Assumption 2.2. In particular, all appearing unknown elements are drawn independently from $U[-1.5, 1.5]$. We then work with the first specification which satisfies the standard assumptions. In this paper, we consider the following cases:

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

For model 1, we use the following specification:

$$\Delta Y_t = \alpha\beta' Y_{t-1} + B_1 \Delta Y_{t-1} + u_t \tag{16}$$

with

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

and

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

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

For the simple three dimensional model 1, rank and lag selection results are reported in Table 1. The results indicate that lag selection performs well independently of the exact choice of tuning parameters with almost perfect results. For rank selection in this simplest case, the penalty term should not be too large i.e. we require $c = 1$ with $\gamma = 2$ for good finite-sample performance. Models 2 and 3 are both of dimension $m = 8$, where

Model 1 ($T = 200$, $\rho = 0.0$)			
	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	$\underline{100}/\underline{95}$	$100/100$	$96/100$
$\gamma = 3.0$	$98/100$	$80/100$	$59/100$
$\gamma = 4.0$	$80/100$	$50/100$	$24/100$
$\gamma = 5.0$	$57/100$	$22/99$	$10/98$

Model 1 ($T = 500$, $\rho = 0.0$)			
	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	$\underline{100}/\underline{99}$	$100/100$	$100/100$
$\gamma = 3.0$	$100/100$	$100/100$	$99/100$
$\gamma = 4.0$	$100/100$	$87/100$	$62/100$
$\gamma = 5.0$	$88/100$	$50/100$	$20/100$

Model 1 ($T = 200$, $\rho = 0.6$)			
	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	$\underline{100}/\underline{86}$	$100/100$	$92/100$
$\gamma = 3.0$	$98/100$	$80/100$	$58/100$
$\gamma = 4.0$	$79/100$	$48/100$	$27/100$
$\gamma = 5.0$	$54/100$	$27/100$	$14/100$

Model 1 ($T = 500$, $\rho = 0.6$)			
	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	$\underline{100}/\underline{81}$	$100/99$	$100/100$
$\gamma = 3.0$	$100/100$	$100/100$	$97/100$
$\gamma = 4.0$	$98/100$	$89/100$	$66/100$
$\gamma = 5.0$	$89/100$	$55/100$	$28/100$

Table 1: Absolute numbers XX/YY of correct model selections by solving (11) and (12) for $b = 100$ repetitions of model 1 with $m = 3$, $r = 2$, $p = 1$. For each parameter specification, XX denotes the number of correct rank selections while YY is the number of correct lag length identifications. Underlining marks the choice with tuning parameters selected according to BIC.

traditional methods cannot be employed either due to inconsistency in theory or because of numerical inefficiency. The selection results for model 2 with $p = 1$ with $r = 4$ in Table 2 demonstrate perfect performance in rank and lag selection generally for a wide range of tuning parameters with $c \geq 1$ and $\gamma \geq 3$. This also holds even for the most difficult case with $\rho = 0.6$ and $T = 200$, while for all other settings the range of acceptable parameters is even wider. In comparison to the simple model 1, larger tuning parameters are preferred both for rank and lag selection due to the higher complexity of the true model. Note that in all cases, the results are based on a ridge regression pre-estimate (13) for the lag choice criterion (12) in order to handle multicollinearity effects. Lag selection results based on adaptive weights from least squares pre-estimates perform substantially inferior.[5] The increased lag length $p = 2$ with $r = 2$ poses the challenge in model 3. There, in particular in the case of 200 observations, larger tuning parameters are preferred for rank selection.

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

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

	$c = 1$	$c = 2$	$c = 3$
$\gamma = 2.0$	99/34	100/72	99/84
$\gamma = 3.0$	100/<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 2: Absolute numbers of correct rank/lag selections by solving (11) and (12) for $b = 100$ repetitions for model 2 and 3 with $m = 8,\ r = 2,\ p = 2$. Reporting style is as in Table 1.

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

Table 3: Absolute numbers of correct rank/lag selections by solving (11) and (12) for $b = 100$ repetitions for model 4 with $m = 16,\ r = 8,\ p = 1$. Reporting style is as in Table 1.

For model 4, we consider a nonstationary VAR(2) process like in model 1 but of dimension 16, i.e. $m = 16$, $r = 8$ and $p = 1$. Due to the complexity from the higher dimensionality of the model we only report results for $T = 500$. For well-chosen tuning parameters, both rank and lag selection results are perfect. In particular, $\gamma = 2$ with larger c and $\gamma = 3$ with smaller c are crucial for good performance of rank selection. Given the complexity of the model, however, there is still a range of such admissible tuning parameters which ensures robust performance in application scenarios where tuning parameters must be pre-chosen. As for models 2 and 3, we use a ridge regression estimate for $\check{B}$ in the lag selection criterion (12). 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.

For known true model specifications, we estimate all four models above according to the refined Lasso procedure (14) and compare estimation fits and one-step ahead forecasts to reduced rank regression. For the case of model 1, we also illustrate their finite-sample advantage if the model is known to the adaptive Lasso estimates from the model selection procedure. In particular, we use $\hat{\Pi}_{adaptive} = \hat{R}_r' \widetilde{S}_r'$ where $\hat{R}_r'$ comprises the first r columns of the solution to the adaptive Lasso rank selection problem (11) and $\widetilde{S}_r'$ consists of the first r rows of the orthonormal matrix defined in (4). 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 $\Sigma_u^{-\frac{1}{2}}$.

$T = 200$	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 4: Comparison of different estimation methods for Model 1

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 5). Similar results are reported in Table 6 for model 3 and Table 7 for model 4. While in the standard low-dimensional model 1, the advantage of using Lasso is not so significant, we find that the more complicated the model is, the more superior becomes the Lasso in particular in estimation. Moreover, the obtained simulation results confirms the advantage of element-wise penalization on the loading matrix over penalization on 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.

6 Empirical Results

Our empirical analysis uses quarterly log FX-rates from Engel et al. (2015) who provide detailed descriptions of the data and their sources.

$T = 200$	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 5: 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 6: Comparison of different estimation methods for Model 3

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

Table 7: Comparison of different estimation methods for Model 4

The bilateral exchange rates are end of quarter values of U.S. dollar (USD) v.s. 17 OECD countries: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Japan, Italy, Korea, Netherlands, Norway, Spain, Sweden, Switzerland, and the United Kingdom. The period under observation runs from the first quarter of 1973 to fourth quarter of 2007 for a total of 140 observations. For an overview of the employed series see Figure 1. Before the cointegration analysis can be carried out, we first control for the statistical requirements for the existence of a cointegration relationship. In particular, we apply panel unit root tests to both original y_{it} and differenced data Δy_{it}, with the corresponding p-values reported in Table 8. The results indicate clearly that the differenced data are stationary while y_{it} are not.

Our results in Figure 2 clearly show the superiority of our technique with respect to a simple random walk for selected examples. For results of all other countries please see the plots in the appendix Figure 3 and 4. All plots highlight that for the case of FX-rates, the tailored lasso is crucial for obtaining not only precise rank and lag estimates but also a column-wise precise estimate of the model specification. In particular, when using the general high-dimensional groupwise Lasso approach of Liang and Schienle (2019), direct estimates for the elements of the cointegration vector are too imprecise, such that the prediction advantage in Figure 2 does not prevail. Though results can be approximately recovered as a robustness check when selecting the cointegration rank and lag structure with the general groupwise lasso first and then keeping only those elements as non-zero which are determined as significant from elementwise significance tests.

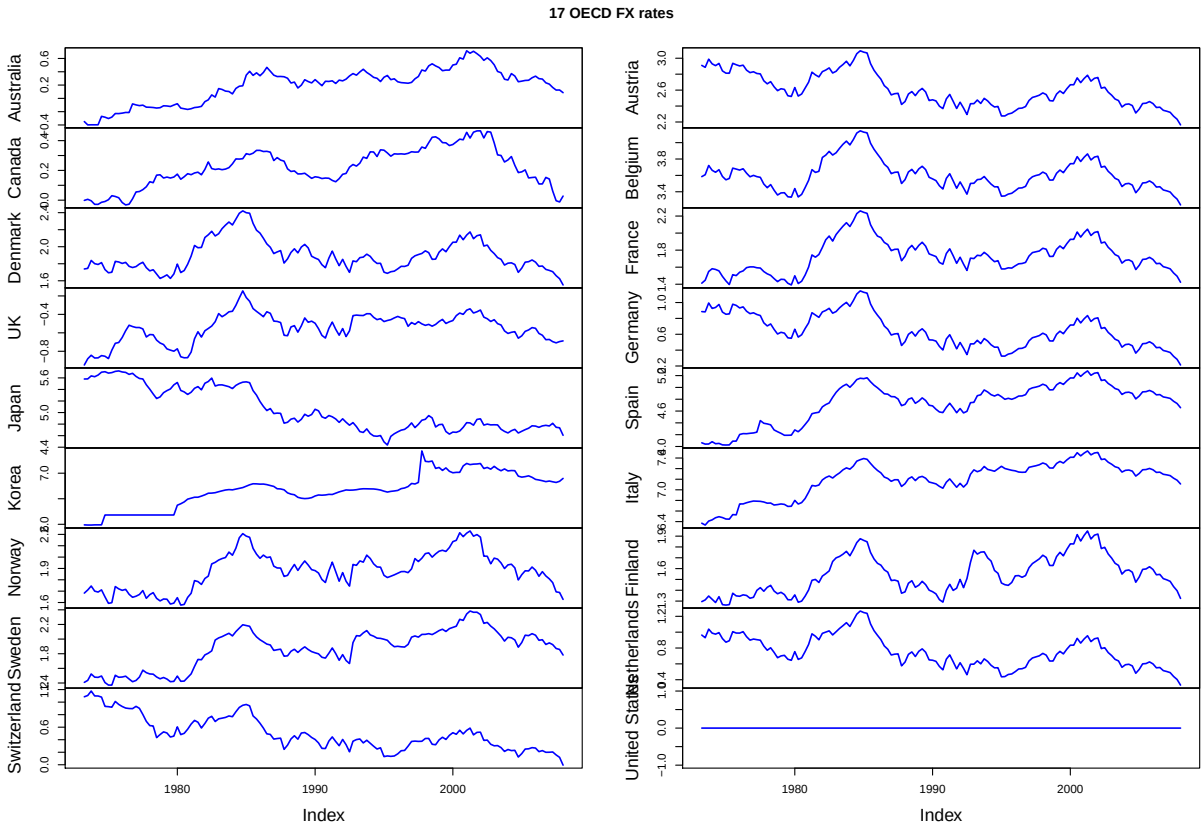

Figure 1: The graph depicts all analyzed FX rates for the 17 OECD countries.

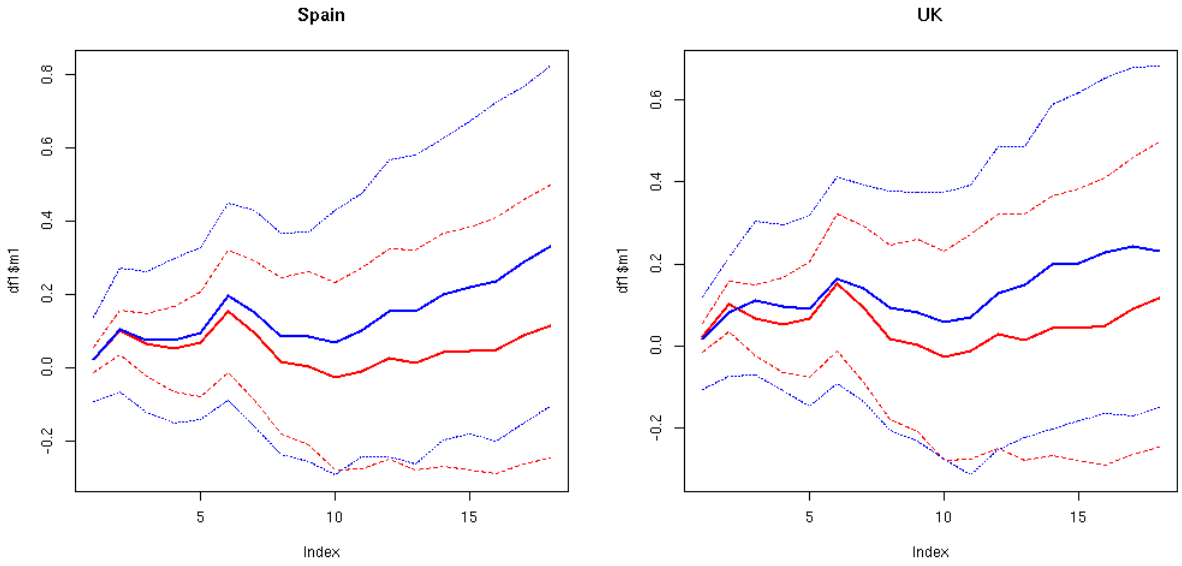

Figure 2: Both figures depict the h-step ahead forecast errors (solid lines) for varying h from our VECM model in red and a simple random walk benchmark in blue. For each h, the interval between the two dotted lines marks the respective bootstrapped pointwise 95% prediction accuracy interval in each case. Results for the FX rate of Spain are reported on the left, results for the UK on the right.

	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 8: The panel unit root tests of foreign exchange time series for each country.

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.

Andrew, G. and Gao, J. (2007). Scalable training of l 1-regularized log-linear models. In *Proceedings of the 24th international conference on Machine learning*, pages 33–40. ACM.

Boswijk, H. P., Jansson, M., and Nielsen, M. O. (2012). Improved likelihood ratio tests for cointegration rank in the var model. Tinbergen Institute Discussion Paper 12-097/III, Amsterdam and Rotterdam.

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.

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

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

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.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *ArXiv e-print*, 1001.0736.

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

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2-3):231 – 254.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica*, 59(6):pp. 1551–1580.

Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):pp. 1356–1378.

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

Koh, K., Kim, S.-J., Boyd, S., and Lin, Y. (2007). An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine Learning Research*, 2007.

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.

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.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):pp. 267–288.

Toda, H. Y. and Phillips, P. C. (1993). Vector autoregressions and causality. *Econometrica: Journal of the Econometric Society*, pages 1367–1393.

Wang, H. and Leng, C. (2008). A note on adaptive group lasso. *Computational Statistics & Data Analysis*, 52(12):5277–5286.

Wilms, I. and Croux, C. (2016). Forecasting Using Sparse cointegration. *International Journal of Forecasting*, 32:1256âĂŞ1267.

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. (2015). Identifying Cointegration by Eigenanalysis. *ArXiv e-print*, 1505.00821.

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

Proof of Lemma 2.1

Define $D_T = diag(\sqrt{T}I_r, TI_{m-r})$ and $E_Q = Q(\widetilde{\Pi} - \Pi)Q^{-1}D_T$. Asymptotically each element in matrix E_Q is finite in probability.

Define an orthonormal matrix $\Xi = \begin{bmatrix} \beta' \\ \beta'_\perp \end{bmatrix}$, then $\Xi\Pi' = \begin{bmatrix} \alpha' \\ 0 \end{bmatrix}$, and

$$\begin{aligned} \Xi\widetilde{\Pi}' &= \Xi\Pi' + \Xi Q' D_T^{-1} E_Q' Q^{-1\prime} \\ &= \begin{bmatrix} \alpha' \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{1}{\sqrt{T}} I_r & \frac{1}{T}\beta'\alpha_\perp \\ 0 & \frac{1}{T}\beta'_\perp\alpha_\perp \end{bmatrix} E_Q' Q^{-1\prime} \end{aligned} \tag{17}$$

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

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

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

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

Thus

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

By the last equality of (17), $\beta'_\perp\widetilde{S}_1\widetilde{R}_1 + \beta'_\perp\widetilde{S}_2\widetilde{R}_2$ satisfies $O_p(\frac{1}{T})$ element-wise. Thus we conclude each element in $\beta'_\perp\widetilde{S}_1$ has order $O_p(\frac{1}{T})$ due to the fact that $\widetilde{R}_1$ has full row rank and $\widetilde{S}_1$ is orthogonal to $\widetilde{S}_2$. Thus $||\beta'_\perp\widetilde{S}_1||_F = O_p(\frac{1}{T})$, which means that the subspace generated by $\widetilde{S}_1$ is a consistent estimate for that generated by β. Moreover, since $\Xi\widetilde{S}_1$ is an orthonormal matrix, $\beta'\widetilde{S}_1$ converges to an orthonormal one, denoted by H at the rate of T. Mathematically, by

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

and

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

the following can be derived

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

or equivalently in finite dimensional case,

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

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

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

If $\|\beta'(\widetilde{S}_1 - \beta H)\|_F$ converges to zero asymptotically, we have either $(\widetilde{S}_1 - \beta H) \in S(\beta_{\perp})$ or $\|\widetilde{S}_1 - \beta H\|_F \to 0$. The first possiblity is excluded by $||\beta'_{\perp}\widetilde{S}_1||_F \to 0$, Therefore, we conclude that asymptotically, $\widetilde{S}_1$ and β characterize the same space with equivalent matrix representations.
Lastly, due to the faster rate of convergence for β, the asymptotic distribution of the estimate for α is not affected by the finite-sample error. By the sparse structure of $\widetilde{R}$ imposed by QR-decomposition, the asymptotic distribution depends on the relevant part only, which is similar to that of adaptive Lasso.

□

Lemma A.1. *With the notation defined in Section 2.2, we have*

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

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

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

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 2.2

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

$$\begin{aligned} & Q(\widetilde{\Pi} - \Pi)Q^{-1}D_T \\ = \; & QUMY'_{-1}Q'D_T^{-1}(D_T^{-1}QY_{-1}MY_T^{-1}Q'D_T^{-1})^{-1} \\ = \; & QUMZ'_{-1}D_T^{-1}(D_T^{-1}Z_{-1}MZ'_{-1}D_T^{-1})^{-1} \end{aligned}$$

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

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

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

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

We have

$$\Delta Z_t = Q\Pi Q^{-1}Z_{t-1} + \psi_t \tag{18}$$

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

$$\psi_t = \Theta(L)v_t \tag{19}$$

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 2.3

The proof directly follows from Lemma 2.2 and Lemma 2.1. □

Proof of Theorem 2.1

For a general form like $y = X\beta + u$, where X has dimension $n \times p$, $\frac{1}{n}X'X$ has full rank and converges to Σ in probability. The solution to ridge regression, i.e., $\arg\min_\beta ||y - X\beta||^2 + v||\beta||_1$, is $\beta_R = (X'X + \nu I_p)^{-1}X'y$. Therefore, $\sqrt{n}(\beta_R - \beta) = -(\frac{1}{n}X'X + \frac{\nu}{n}I_p)^{-1}\frac{\nu}{\sqrt{n}}\beta + (\frac{1}{n}X'X + \frac{\nu}{n}I_p)^{-1}\frac{1}{\sqrt{n}}X'u$. The bias term $-(\frac{1}{n}X'X + \frac{\nu}{n}I_p)^{-1}\frac{\nu}{\sqrt{n}}\beta \to_p 0$ if $\frac{\nu}{\sqrt{n}} \to_p 0$. Therefore $\lim_{T\to\infty}\widetilde{B}_R = B$ holds. □

Proof of Theorem 3.1

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

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

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

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

$$\begin{aligned}\Delta_T(E_R) &= vec(E_R D_T^{-1})'(\widetilde{S}' Y_{-1} \otimes I_m)(I_T \otimes \Sigma_u^{-1})(Y_{-1}'\widetilde{S} \otimes I_m) vec(E_R D_T^{-1}) \\ &- 2vec(U)'(I_T \otimes \Sigma_u^{-1})(Y_{-1}'\widetilde{S} \otimes I_m) vec(E_R D_T^{-1}) \\ &+ \sum_{i,j=1}^{m} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^\gamma} (|R(i,j) + E_R D_T^{-1}(i,j)| - |R(i,j)|) \\ &= vec(E_R)'(D_T^{-1}\widetilde{S}' Y_{-1} \otimes I_m)(I_T \otimes \Sigma_u^{-1})(Y_{-1}'\widetilde{S} D_T^{-1} \otimes I_m) vec(E_R) \\ &- 2vec(\Sigma_u^{-1} U Y_{-1}'\widetilde{S} D_T^{-1})' vec(E_R) \\ &+ \sum_{i,j=1}^{m} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^\gamma} (|R(i,j) + E_R D_T^{-1}(i,j)| - |R(i,j)|) \qquad (20) \\ &= vec(E_R)'(D_T^{-1}\widetilde{S}' \sum_{t=1}^{T} Y_{t-1} Y_{t-1}'\widetilde{S} D_T^{-1} \otimes \Sigma_u^{-1}) vec(E_R) \\ &- 2vec(\sum_{t=1}^{T} \Sigma_u^{-1} u_t Y_{t-1}'\widetilde{S} D_T^{-1})' vec(E_R) \\ &+ \sum_{i,j=1}^{m} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^\gamma} (|R(i,j) + E_R D_T^{-1}(i,j)| - |R(i,j)|)\end{aligned}$$

In Lemma 2.1 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

$$\begin{aligned} & D_T^{-1}\widetilde{S}'\sum_{t=1}^{T} Y_{t-1}Y_{t-1}'\widetilde{S}D_T^{-1} \\ = & \begin{bmatrix} I_r + O_p(\frac{1}{T}) & \sqrt{T}O_p(\frac{1}{T}) \\ \frac{1}{\sqrt{T}}\widetilde{S}_2'q_1 & \widetilde{S}_2'q_2 \end{bmatrix} \begin{pmatrix} T^{-1}\sum_{t=1}^{T} Z_{1,t-1}Z_{1,t-1}' & T^{-3/2}\sum_{t=1}^{T} Z_{1,t-1}Z_{2,t-1}' \\ T^{-3/2}\sum_{t=1}^{T} Z_{2,t-1}Z_{1,t-1}' & T^{-2}\sum_{t=1}^{T} Z_{2,t-1}Z_{2,t-1}' \end{pmatrix} \\ & \begin{bmatrix} I_r + O_p(\frac{1}{T}) & \frac{1}{\sqrt{T}}q_1'\widetilde{S}_2 \\ \sqrt{T}O_p(\frac{1}{T}) & q_2'\widetilde{S}_2 \end{bmatrix} \qquad (21) \\ \to_d & \begin{bmatrix} \Sigma_{z1z1} & 0 \\ 0 & \widetilde{S}_2'q_2\left(\left([0 \quad I_{m-r}]\Sigma_v^{1/2}(\int_0^1 W_m W_m' ds)\Sigma_v^{1/2}\begin{bmatrix} 0 \\ I_{m-r} \end{bmatrix}\right)^{-1}\right)q_2'\widetilde{S}_2 \end{bmatrix} \end{aligned}$$

For the second term in equation (20), we have:

$$\begin{aligned} & vec(\Sigma_u^{-1}(\sum_{t=1}^{T} u_t Y_{t-1}')\widetilde{S}D_T^{-1}) = vec(\Sigma_u^{-1}(\sum_{t=1}^{T} u_t Y_{t-1}'Q'D_T^{-1})D_T Q'^{-1}\widetilde{S}D_T^{-1}) \\ = & vec(\begin{bmatrix} T^{-1/2}\sum \Sigma_u^{-1}u_t Z_{1,t-1}' & T^{-1}\sum \Sigma_u^{-1}u_t Z_{2,t-1}' \end{bmatrix} \begin{bmatrix} I_r + O_p(\frac{1}{T}) & \frac{1}{\sqrt{T}}q_1'\widetilde{S}_2 \\ \sqrt{T}O_p(\frac{1}{T}) & q_2'\widetilde{S}_2 \end{bmatrix}) \\ \to_d & \begin{bmatrix} N(0, \Sigma_{z1z1}\otimes\Sigma_u^{-1}) \\ vec\{\Sigma_u^{-1}Q^{-1}\Sigma_v^{\frac{1}{2}}(\int_0^1 W_m dW_m')'\Sigma_v^{\frac{1}{2}}\begin{bmatrix} 0 \\ I_{m-r} \end{bmatrix}q_2'\widetilde{S}_2\} \end{bmatrix} \qquad (22) \end{aligned}$$

Next we should pay attention to the last term in eq. (20).
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

$$X_{k'}'(I_T \otimes \Sigma_u^{-1})(vec(\Delta Y) - X vec(\hat{R}_T')) = \frac{1}{2} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^{\gamma}} sign(\hat{R}_T'(i,j)) \tag{23}$$

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

$$\frac{1}{T_{k'}} X_{k'}'(I_T \otimes \Sigma_u^{-1})(vec(\Delta Y) - X vec(\hat{R}_T')) = \frac{1}{T_{k'}} \frac{1}{2} \frac{\lambda_{i,j,T}^{rank}}{|\widetilde{R}(i,j)|^{\gamma}} sign(\hat{R}_T'(i,j)) \tag{24}$$

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 (24) 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 3.2

The proof directly follows from Theorem 3.1 and Lemma 2.3 □

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

□

Distribution of $\widetilde{\Pi}$ under Assumption 4.1

Lemma A.2. *If error terms u_t in equation (2) are defined in Assumption 4.1, then the least squares estimate for Π is distributed as*

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

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

When the error terms are dependent, the stochastic part $\{u_t Z_{1,t-1}'\}$ 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,

$$\frac{1}{T}\sum_{t=1}^{T} Qu_t Z'_{1,t-1} \quad = \quad \frac{1}{T}\sum_{t=1}^{T} Qw_t Z'_{1,t-1} + \frac{1}{T}\sum_{t=1}^{T} Q(\kappa(L)-\kappa(0))w_t Z'_{1,t-1}$$

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

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

After being corrected for the bias term, the asymptotic distribution has similar form with the $i.i.d$ error case. The asymptotics of the unit root process under Assumption 4.1 can be referred to Lütkepohl (2007)

□

Proof of Proposition 4.1

The proof is similar to the proof of Theorem 3.1 except that the coefficient matrix R is from the QR decomposition of $\Pi + \Upsilon\Sigma_{z1}^{-1}\beta'$, the biased counterpart. The argument with respect to the penalty should be modified as follows.
If at least one element in $R(i,)$ is non-zero, then

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

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

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

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

Proof of Theorem 4.1

As in the proof of Theorem 3.1, we define such an objective function:

$$\begin{aligned} \Psi_T(E) = & \left\| vec(\Delta Y) - (\left[\begin{array}{cc} Y_{-1}'\hat{\beta}^{\dagger} & \Delta X^{p\prime} \end{array}\right] \otimes I_m) vec(\left[\begin{array}{cc} \alpha & B^p \end{array}\right] + \frac{1}{\sqrt{T}}E) \right\|_{I_T \otimes \Sigma_u^{-1}}^2 \\ + & \sum_{i=1}^{m}\sum_{j=1}^{r}\lambda_{i,j,T}^{rank}|\alpha(i,j)+\frac{1}{\sqrt{T}}E_0(i,j)| \\ + & \sum_{k=1}^{p}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_{i,j,T}^{lag,k}|B_k(i,j)+\frac{1}{\sqrt{T}}E_k(i,j)| \end{aligned} \tag{25}$$

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

$$\begin{aligned} \Delta_T(E) = & \Psi_T(E) - \Psi_T(0) \\ = & vec(\frac{1}{\sqrt{T}}E)'(\left[\begin{array}{c} \hat{\beta}^{\dagger\prime}Y_{-1} \\ \Delta X^p \end{array}\right] \otimes I_m)(I_T \otimes \Sigma_u^{-1})(\left[\begin{array}{cc} Y_{-1}'\hat{\beta}^{\dagger} & \Delta X^{p\prime} \end{array}\right] \otimes I_m) vec(\frac{1}{\sqrt{T}}E) \\ - & 2vec(U)'(I_T \otimes \Sigma_u^{-1})(\left[\begin{array}{cc} Y_{-1}'\hat{\beta} & \Delta X^{p\prime} \end{array}\right] \otimes I_m) vec(\frac{1}{\sqrt{T}}E) \\ + & \sum_{i=1}^{m}\sum_{j=1}^{r}\lambda_{i,j,T}^{rank}(|\alpha(i,j)+\frac{1}{\sqrt{T}}E_0(i,j)| - |\alpha(i,j)|) \\ + & \sum_{k=1}^{p}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_{i,j,T}^{lag,k}(|B_k(i,j)+\frac{1}{\sqrt{T}}E_k(i,j)| - |B_k(i,j)|) \end{aligned} \tag{26}$$

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:

$$\frac{1}{T}\begin{bmatrix} \hat{\beta}^{\dagger}Y_{-1} \\ \Delta X^{p} \end{bmatrix}\begin{bmatrix} Y'_{-1}\hat{\beta}^{\dagger} & \Delta X^{p\prime} \end{bmatrix} \rightarrow_{p} \Sigma_{\Gamma^{p}\Gamma^{p}}$$

Based on the proof of Theorem 3.1, we can similarly show that

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

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, $\frac{\lambda_{i,j,T}^{rank}}{\sqrt{T}} \rightarrow 0$. 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 4.1 follows naturally.

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

□

B Additional Results

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

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

$$T(\hat{\beta}_2\hat{\beta}_1^{-1} - \beta_0) \to_d (\int_0^1 W_{m-r}^* dW_r^*)'(\int_0^1 W_{m-r}^* W_{m-r}^{*'} ds)^{-1} \tag{27}$$

where

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

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 $\Sigma_v^{1/2}(\int_0^1 W_m dW_m')'\Sigma_v^{1/2} \left[\begin{array}{c} 0_{r\times(m-r)} \\ I_{m-r} \end{array} \right]$ 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 Specifications for Simulations

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

$$\alpha = \left[\begin{array}{rrrr} -1.47 & -1.3 & 0 & -1.26 \\ 0 & 0.97 & 0 & 0 \\ 0 & 0 & -0.74 & 0 \\ -1.19 & 0.85 & 0 & 0 \\ -0.55 & 0.78 & -1 & -1.37 \\ 0.8 & 0.75 & 0 & 0 \\ 0 & -0.74 & -1.26 & -0.78 \\ 0 & -1.4 & 0 & 0 \end{array} \right]$$

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

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}$$

D Additional Empirical Results

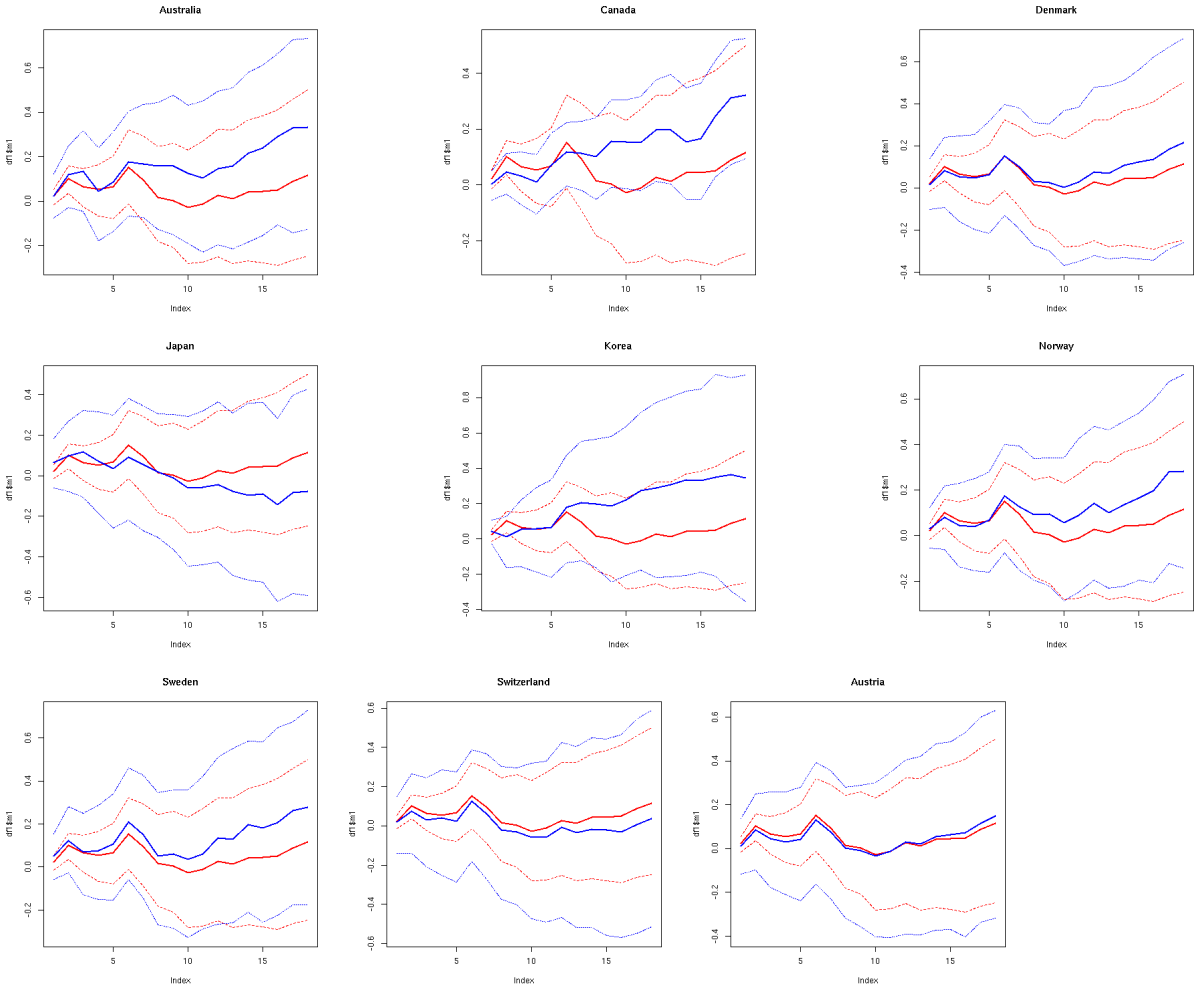

Figure 3: All figures depict the h-step ahead forecast errors (solid lines) for varying h from our VECM model in red and a simple random walk benchmark in blue. For each h, the interval between the two dotted lines marks the respective bootstrapped pointwise 95% prediction accuracy interval in each case.

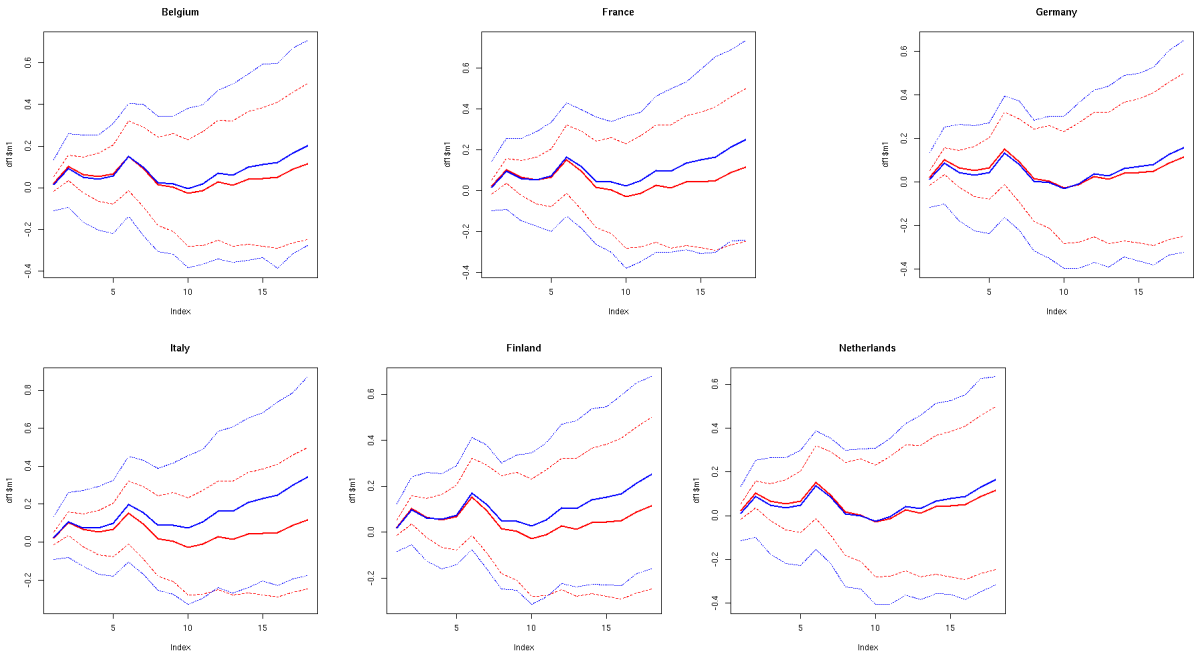

Figure 4: All figures depict the h-step ahead forecast errors (solid lines) for varying h from our VECM model in red and a simple random walk benchmark in blue. For each h, the interval between the two dotted lines marks the respective bootstrapped pointwise 95% prediction accuracy interval in each case.