Journal of Econometrics 208 (2019) 418–441

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Determination of vector error correction models in high dimensions☆

Chong Liang, Melanie Schienle *

Karlsruhe Institute of Technology (KIT), Germany

ARTICLE INFO

Article history:
Received 1 September 2015
Received in revised form 16 August 2018
Accepted 10 September 2018
Available online 5 November 2018

JEL classification:
C32
C52

Keywords:
High-dimensional time series
VECM
Cointegration rank and lag selection
Lasso
Credit default swap

ABSTRACT

We provide a shrinkage type methodology which allows for simultaneous model selection and estimation of vector error correction models (VECM) when the dimension is large and can increase with sample size. Model determination is treated as a joint selection problem of cointegrating rank and autoregressive lags under respective practically valid sparsity assumptions. We show consistency of the selection mechanism by the resulting Lasso-VECM estimator under very general assumptions on dimension, rank and error terms. Moreover, with computational complexity of a linear programming problem only, the procedure remains computationally tractable in high dimensions. We demonstrate the effectiveness of the proposed approach by a simulation study and an empirical application to recent CDS data after the financial crisis.

1. Introduction

Complex financial systems are dynamic, high-dimensional and often contain a large number of non-stationary potentially cointegrated components. Examples include the degree of interdependence of corporate debt among different banks and its interplay with sovereign debt, both measured as a large system of credit default spreads (CDS), but also risk analysis for large-dimensional portfolios containing many different nonstationary elements such as exchange or interest rates in the presence of a limited number of applicable observations during crisis times. Generally, the standard tool to handle multivariate nonstationary time-series has been the vector error correction model (VECM) as introduced by Engle and Granger (1987). But even for settings with a fixed number of dimensions mildly exceeding two, existing econometric techniques for VECM often fail to provide accurate, testable and computationally tractable estimates, e.g. sequential Johansen tests (Johansen, 1988, 1991) and refinements thereof such as e.g. Xiao and Phillips (1999), Hubrich et al. (2001), Boswijk et al. (2012), Cavaliere et al. (2012). In the high-dimensional case of the examples before, however, also information criteria based techniques such as Chao and Phillips (1999) are no longer applicable and novel types of methods are required which need a completely different statistical analysis. Such techniques are important for understanding the explicit interplay of different market components in order to judge their systemic importance and market relevance.

☆ We are grateful to Weibiao Wu, Qiwei Yao, Rongmao Zhang, Kyusang Yu and Bernd Droge for helpful discussions. We furthermore thank the editor, Oliver Linton, and three anonymous referees for constructive comments which led to substantial improvements of the paper. We acknowledge support from Deutsche Forschungsgemeinschaft through grant SCHI-1127.

* Correspondence to: Karlsruhe Institute of Technology, Department of Economics (ECON), Chair of Econometrics and Statistics, Blücherstr. 17, 76185, Karlsruhe, Germany.
E-mail address: melanie.schienle@kit.edu (M. Schienle).

https://doi.org/10.1016/j.jeconom.2018.09.018
0304-4076/

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

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

Our proposed technique is particularly related to a recent literature which uses Lasso in a high-dimensional time series context. Kock and Callot (2015) and Basu and Michailidis (2015) provide model determination techniques in a stationary high-dimensional VAR context. There has also been a recent empirical literature which employs Lasso-type penalizing algorithms for VECM without mathematical proofs, see Signoretto and Suykens (2012), Wilms and Croux (2016). To the best of our knowledge, comparable settings of determining cointegrated time series have only been investigated in three recent theoretical papers by Liao and Phillips (2015) in fixed dimensions and Zhang et al. (2018) and Onatski and Wang (2018) in high dimensions. In particular for fixed dimensions, Liao and Phillips (2015) are the first to propose a Lasso-procedure for VECM with theoretical proofs. Their procedure, however, penalizes the eigenvalues of a generally asymmetric matrix which limits the applicability of the technique to specific fixed dimensional settings. Zhang et al. (2018) provide statistical results for a factor model dealing with high-dimensional non-stationary time series with a focus on forecasting without employing a VECM structure. The focus of Onatski and Wang (2018) is not on model selection consistency but on asymptotic distributions of the eigenvalues.

The rest of the paper is organized as follows. In Sections 2 and 3, we derive the Lasso objective function in a VECM specification in order to determine the cointegration rank and the VAR lags. The consistency results will be derived. Section 4 extends the previous econometric analysis to a more general setting with non *i.i.d.* innovations. In Section 5 we study the finite-sample performance of the method in several simulation experiments. We also provide an empirical application to CDS data for European countries and banks in Section 6. Section 7 concludes. Proofs for Sections 2 and 3 are contained in the Appendix . Proofs for Section 4 and technical Lemma are in the online supplementary.

Throughout the paper, we use the following notation. For a vector $x \in R^m$, the l_2 norm is defined as $\|x\|_2 = \sqrt{\sum_{j=1}^m x_j^2}$ and $\|x\|_\infty = \sup_{1\leqslant j\leqslant m} |x_j|$ is the l_∞ norm. For a matrix $A = ((A_{ij}))$ of dimension $m \times l$, $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^l A_{ij}^2}$ denotes the Frobenius norm and $\|A\|_2 = \sup\{\|Ax\|_2 : x \in R^l \text{ with } \|x\|_2 = 1\}$ the l_2 norm. Besides, we denote by $\lambda_j(C)$ the jth largest eigenvalue of a square matrix C in absolute value, where as $\sigma_j(A)$ is the j-largest singular value of A, i.e. $\sigma_j^2(A) = \lambda_j(A'A)$. Without loss of generality, we assume the singular values to be non-negative for notational convenience. We use $vec(A) = [A'_{\cdot 1}, A'_{\cdot 2}, \ldots, A'_{\cdot n}]'$ for vectorizing a matrix A by stacking all columns where $A_{\cdot j}$ is the jth column in matrix A. For $rank(A) = l < m$, the orthogonal complement to a matrix A is defined as $\mathcal{A}_\perp = \{U \in \mathbb{R}^{m\times(m-l)} | U'A = 0\}$. For an orthonormal $A_\perp$ of A it holds that $A_\perp \in \mathcal{A}_\perp$ and in addition that $A'_\perp A_\perp = I_{m-l}$.

2. Cointegration rank selection

2.1. Set-up and fundamental results

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

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

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

for $t = 1, \ldots, T$, where B_k are $m \times m$ stationary lag coefficient matrices for $k = 1, \ldots, P$ and Π is the $m \times m$ cointegration matrix of rank r with $0 \leqslant r < m$ marking the number of cointegration relations in the system. In case of $r = m$, all the

components in Y_t are stationary, which is not relevant to our non-stationary time series setting. Π can be decomposed as $\Pi = \alpha\beta'$, where $\beta \in \mathbb{R}^{m\times r}$ constitutes the r long-run cointegrating relations and $\alpha \in \mathbb{R}^{m\times r}$ is a loading matrix of rank r. This decomposition is unique up to a nonsingular matrix H, so only the space of cointegration relations is identified but not β. Without loss of generality, we set β as orthogonal, i.e. $\beta'\beta = I_r$.

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

In the following, we work with the matrix version of (1)

$$\Delta Y = \Pi Y_{-1} + B\Delta X + W \tag{2}$$

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

For model selection, we disentangle the joint lag-rank selection problem by employing a Frisch–Waugh-idea in the VECM model (2). With this, we obtain two independent criteria for lag and rank choice which can be computed separately. For rank selection, the partial least squares pre-estimate $\widetilde{\Pi}$ can be obtained from the corresponding partial model when removing the effect of ΔX in ΔY and Y_{-1} by regressing $\Delta Y M_{\Delta x}$ on $Y_{-1}M_{\Delta x}$ with $M_{\Delta x} = M = I_T - \Delta X'(\Delta X \Delta X')^{-1}\Delta X$. Therefore, (2) is equivalent to

$$\Delta\widetilde{Y}_t = \alpha\beta'\widetilde{Y}_{t-1} + \widetilde{w}_t \tag{3}$$

with components $\Delta\widetilde{Y} = \Delta Y M$, $\widetilde{Y}_{-1} = Y_{-1}M$ and $\widetilde{W} = WM$. Thus model selection is reduced to rank selection only in (3).

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

Assumption 2.1. For the error component w_t in (1) exists a representation $w_t = \Sigma_w^{1/2} e_t$ where the elements satisfy the following conditions

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

The requirement of i.i.d. components in the error term representation allows focusing on the key aspects of our Lasso selection procedure in the high dimensional set-up while keeping technical results to a minimum. In Section 4, we show how this Assumption can be generalized admitting linear forms of weak dependence. Such a general setting, however, requires a proof for a general strong invariance principle which is key for our consistency results but not available under weak dependence for high dimensions in the literature so far.

From the first two points in Assumption 2.1, Σ_w denotes the covariance matrix of w_t. The third point imposes a sparse structure and ensures positive definiteness of Σ_w through bounding the smallest eigenvalue of Σ_w away from zero. This sparsity condition is satisfied if Σ_w is a banded diagonal matrix with off-diagonal entries far away from the diagonal decaying to zero fast enough (see e.g. Bickel and Levina (2008)). In practice, this seems plausible e.g. in the case of sovereign CDS as treated in the empirical example that geographical distance between countries implies such a cross-section decay structure in the innovations naturally.

Our shrinkage selection procedure for the cointegration rank is based on a least squares pre-estimate of Π from the $M_{\Delta x}$-transformed VECM equation (3)

$$\widetilde{\Pi} = \left(\Delta Y M Y'_{-1}\right)\left(Y_{-1}MY'_{-1}\right)^{-1} \tag{4}$$

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

Assumption 2.2.

1. The roots for $|(1-z)I_m - \Pi z - \sum_{j=1}^p B_j(1-z)z^j| = 0$ are either $|z| = 1$ or $|z| > 1$.
2. The number of roots lying on the unit circle is $m - r$.
3. The matrix $\alpha'_\perp(I_m - \sum_{i=1}^p B_i)\beta_\perp$ is nonsingular with $\|(\alpha'_\perp(I_m - \sum_{i=1}^p B_i)\beta_\perp)^{-1}\|_2 < \infty$.

The last point of Assumption 2.2 is a stronger version than in fixed dimensional case which requires that the smallest singular value of $\alpha_\perp'\beta_\perp$ to be significantly different from zero, which is equivalent to that the basis generating β cannot be close to any of the basis of $\alpha_\perp$.

It is well known that for the standard low-dimensional setup with fixed m in (1) and Assumptions 2.1 and 2.2, the standard least squares estimator in (4) is consistent (see e.g. Lütkepohl, 2007). In our high-dimensional case, however, we need to explicitly derive its statistical properties. These are key for the construction and validity of a Lasso cointegration rank selection procedure in this paper.

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

Assumption 2.3.

1. All singular values $\sigma_j(\alpha)$ of α fulfill $0 < \sigma_r(\alpha) \leqslant \cdots \leqslant \sigma_1(\alpha) < \infty$ and there exist $\tau_1 > 0$ and $K_1 > 0$ such that
$$r^{\tau_1}\sigma_r(\alpha) \geqslant K_1 \ .$$
2. For $B_p = (B_p(i,j))_{i,j=1}^m$ it holds that $\max_{1\leqslant i,j\leqslant m}|B_p(i,j)| \geqslant \varepsilon_B > 0$ with $\varepsilon_B > 0$ and for B defined in (2) there exists a positive $K_B < \infty$ such that $\|B\|_2 < K_B$.

With both dimension m and cointegration rank r increasing with sample size, $\alpha'\alpha$ converges by construction to a compact operator of which the spectrum is well-known to have zero as an accumulation point (cf. Zhao and Yu (2006)). Since therefore the smallest singular value of α in (3) has a converging subsequence to zero, Assumption 2.3 connects the admissible rate of divergence of the rank r with the rate of decay in singular values of α (cf. the high-dimensional factor model literature, e.g. Li et al. (2017)). Thus for deriving statistical properties of corresponding estimates in this set-up this rate that $\sigma_r(\alpha)$ decays to zero restricts the rate at which r can increase with T. We generally denote elements as relevant if they are non-zero in finite samples but with potentially zero limits or accumulation points asymptotically.

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

The statistical properties of $\widetilde{\Pi}$ rely on a Q-transformation of the defining $M_{\Delta x}$-transformed VECM equation (3) which allows to disentangle stationary and nonstationary components. We set $Q = \begin{bmatrix} \beta' \\ \alpha_\perp' \end{bmatrix}$ and $Q^{-1} = \begin{bmatrix} \alpha(\beta'\alpha)^{-1} & \beta_\perp(\alpha_\perp'\beta_\perp)^{-1} \end{bmatrix}$, where $\alpha_\perp$ and $\beta_\perp$ are orthogonal complements of α and β respectively, as defined in Assumption 2.2. After Q-transformation of (3) we get

$$\begin{aligned} \Delta\widetilde{Z}_{1,t} &= \beta'\alpha\widetilde{Z}_{1,t-1} + \widetilde{v}_{1,t} \\ \Delta\widetilde{Z}_{2,t} &= \widetilde{v}_{2,t} \end{aligned} \tag{5}$$

where $\widetilde{Z}_t = Q\widetilde{Y}_t = [(\beta'\widetilde{Y}_t)', (\alpha_\perp'\widetilde{Y}_t)']' = \left[\widetilde{Z}_{1,t}' \widetilde{Z}_{2,t}'\right]'$ and $\widetilde{v}_t = Q\widetilde{w}_t = \left[\widetilde{v}_{1t}' \widetilde{v}_{2t}'\right]'$. Note that by definition, the first component $\widetilde{Z}_{1,t}$ of dimension r is stationary and the $(m-r)$-dimensional remainder $\widetilde{Z}_{2,t}$ is a unit root process. We also denote $Z_t = QY_t = \left[Z_{1,t}' Z_{2,t}'\right]'$, and $v_t = Q\,w_t = \left[v_{1t}' v_{2t}'\right]'$. From (5) the corresponding estimate of the cointegration matrix is obtained as

$$Q\widetilde{\Pi}Q^{-1} = \left(\sum_{t=1}^T \Delta\widetilde{Z}_{t-1}\widetilde{Z}_{1,t-1}' \quad \sum_{t=1}^T \Delta\widetilde{Z}_{t-1}\widetilde{Z}_{2,t-1}'\right) \begin{pmatrix} \sum_{t=1}^T \widetilde{Z}_{1,t-1}\widetilde{Z}_{1,t-1}' & \sum_{t=1}^T \widetilde{Z}_{1,t-1}\widetilde{Z}_{2,t-1}' \\ \sum_{t=1}^T \widetilde{Z}_{2,t-1}\widetilde{Z}_{1,t-1}' & \sum_{t=1}^T \widetilde{Z}_{2,t-1}\widetilde{Z}_{2,t-1}' \end{pmatrix}^{-1} \tag{6}$$

with $\widetilde{\Pi}$ from (4). For this, the statistical properties can be derived in a block-wise way. The result is stated in the following theorem.

Denote by $\mathbf{M} = [\mathbf{M}_1', \mathbf{M}_2']'$ an m-dimensional martingale process with covariance $Q\,\Sigma_w Q'$ with Σ_w from Assumption 2.1 where each component $\mathbf{M}^k$ constitutes a Brownian motion starting at zero and $\mathbf{M}_1$ marks the first subvector of dimension r and $\mathbf{M}_2$ for the vector of the last $m-r$ elements. In the following, given the rank $r < m$, for any matrix $A \in \mathbb{R}^{m\times m}$, denote the top-left $r \times r$ block of A by A_{11}, the bottom-left $(m-r)\times r$ block by A_{12}, the top-right $r \times (m-r)$ block by A_{21}, and the bottom right $(m-r)\times(m-r)$ block by A_{22} respectively.

Theorem 2.1. *Let Assumptions 2.1–2.3 hold. With $D_T = diag(I_r, TI_{m-r})$ define*

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

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

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

$$\|\widetilde{\Psi}_{11} - (\beta'\alpha)\|_F = O_p\left(\frac{r}{\sqrt{T}}\right)$$

$$\|\widetilde{\Psi}_{12} - \mathbf{V}_{12}\|_F = O_p\left(m\sqrt{\frac{(\log T)(\log\log T)^{1/2}}{T^{1/2}}}\right)$$
$$\|\widetilde{\Psi}_{21}\|_F = O_p\left(\sqrt{\frac{mr}{T}}\right)$$
$$\|\widetilde{\Psi}_{22} - \mathbf{V}_{22}\|_F = O_p\left(m\sqrt{\frac{(\log T)(\log\log T)^{1/2}}{T^{1/2}}}\right).$$

Under suitable restrictions on the expansion rates of m and r consistency of all components in $\widetilde{\Psi}$ can be reached. For the stationary components the standard fixed-dimensional $T^{-1/2}$ rate is slowed down by the expansion rates of r and mr. For the nonstationary components, however, the convergence rate depends on the moment conditions of the innovations. In particular, the limit results for the nonstationary blocks in Theorem 2.1 yield stochastic elements of Ψ with a general martingale structure of only elementwise Brownian motions instead of a standard multivariate Brownian motion. This is because generally in the high dimensional set-up, a vector composed of elementwise Brownian motion processes does not necessarily follow a multivariate Brownian motion in contrast to standard multivariate fixed dimensional case, see Kosorok and Ma (2007). With higher moment assumptions on the innovation than Assumption 2.1, however, a Brownian motion type limit and faster rates of convergences could be achieved. Though for general applicability of our subsequent methodology to financial market data, the stated rates are sufficient and we therefore refrain from imposing moments beyond $4+\delta$.

Note that the technical condition $m^{\frac{1}{2\tau_1+1}}$ imposes an upper bound for the expansion rate of the rank r depending on the rate of decay of the smallest singular value $\sigma_r(\alpha)$ in T. Combined with Assumption 2.3, it implies that for fast decreasing subsequences of $\sigma_r(\alpha)$, the polynomial exponent τ_1 must also be large, imposing a binding restriction on the rate of r. Whereas in the case with any subsequence of $\sigma_r(\alpha)$ approaching zero not too rapidly, identification of relevant elements is easier and thus r can increase faster. When the dimension m is fixed, all the singular values of α are significantly different from zero, which is equivalent to assume $\tau_1 = 0$ and thus there is no restriction on r.

We can combine the blockwise results of Theorem 2.1 to obtain the following corollary.

Corollary 2.1. *Let Assumptions 2.1–2.3 hold. Moreover, we require $m = O(T^{1/4-\varepsilon})$ with $\varepsilon \in (0, \frac{1}{4}]$ and $r = O\left(m^{\frac{1}{2\tau_1+1}}\right)$. Then:*

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

with $\widetilde{\Psi}$ as in Theorem 2.1 and $\Psi_0 = Q\Pi Q^{-1} = \begin{pmatrix}\beta'\alpha & 0\\ 0 & 0\end{pmatrix} = E(\Psi)$.

Thus the Q-transformed $\widetilde{\Pi}$ consistently estimates the population counterpart under the stated conditions on m and r. The admissible expansion rate $m = O(T^{1/4-\varepsilon})$ mainly results from the mild $(4+\delta)$ moment condition on the innovations in Assumption 2.1 and the strong invariance principle. Fixed dimensions are included as a special case for $\varepsilon = \frac{1}{4}$. Hence, the relevant r-dimensional stationary part can be consistently identified as all other components of Ψ have expectation 0.

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

The basic principle of standard Lasso-type methods is to determine the number of covariates in a linear model according to a penalized loss-function criterion. Likewise, the determination of the cointegration rank in (1) amounts to distinguishing the vectors spanning the r-dimensional cointegration space from the $(m-r)$ basis of its orthogonal complement. This is also equivalent to separating the r relevant singular values of Π in (3) from the non-relevant ones, where the number of relevant singular values corresponds to the rank. Thus, the corresponding loading matrix for the stationary part $\widetilde{Z}_{1,t} = \beta'\widetilde{Y}_{t-1}$ in (5) is α while the remainder $\beta'_\perp\widetilde{Y}_{t-1}$ should get loading zero in the Q-transformed defining VECM equation (3). We use the QR decomposition with column-pivoting[1] to detect the rank of $\Pi = \alpha\beta' = SR$ as the rank of R, where S is orthonormal, i.e. $S'S = I$, and R is an upper triangular matrix.[2] Column-pivoting orders columns in R according to size putting zero rows at the end.[3] Thus the rank r of Π corresponds to the number of relevant columns in R.

The challenge is, to show that such disentangling of the stationary part $\widetilde{Z}_1$ from the non-stationary $\widetilde{Z}_2$ also works empirically when starting from estimated objects instead of true unobserved population counterparts. Thus calculating the rank from a QR-decomposition with column pivoting of the consistent pre-estimate $\widetilde{\Pi}$ does indeed yield a consistent estimate of the true rank r. In particular, this requires ensuring that true non-relevant singular values, loadings or entries

[1] We denote the orthogonal matrix in the QR-decomposition by S in order to avoid labeling confusion with the Q-transformation used in (5).

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

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

can be distinguished from elements which just appear as non-relevant due to estimation but which in fact truly are relevant which would delude the rank choice. In the following, we show that different speeds of convergence in the stationary and nonstationary parts, however, help to disentangle the two components and can be cleverly exploited in constructing weights for a consistent adaptive group Lasso procedure.

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

$$\widetilde{\Pi}' = \widetilde{R}'\widetilde{S}' = \begin{pmatrix} \widetilde{R}_1' & \widetilde{R}_2' \end{pmatrix} \begin{pmatrix} \widetilde{S}_1' \\ \widetilde{S}_2' \end{pmatrix} = \begin{pmatrix} \widetilde{R}_{11}' & 0 \\ \widetilde{R}_{12}' & \widetilde{R}_{22}' \end{pmatrix} \begin{pmatrix} \widetilde{S}_1' \\ \widetilde{S}_2' \end{pmatrix} \tag{7}$$

where $\widetilde{S}$ is $m \times m$ orthonormal, i.e. $\widetilde{S}'\widetilde{S} = I$, with components $\widetilde{S}_1' \in \mathbb{R}^{r\times m}$ and $\widetilde{S}_2' \in \mathbb{R}^{(m-r)\times m}$. $\widetilde{R}$ is an upper triangular matrix with blocks $\widetilde{R}_1 = (\widetilde{R}_{11} \;\; \widetilde{R}_{12}) \in \mathbb{R}^{r\times m}$ and $\widetilde{R}_2 = (0 \;\; \widetilde{R}_{22}) \in \mathbb{R}^{(m-r)\times m}$ and components with the same notation as for Theorem 2.1 where $\widetilde{R}_{11} \in \mathbb{R}^{r\times r}$, $\widetilde{R}_{12} \in \mathbb{R}^{r\times(m-r)}$, and $\widetilde{R}_{22} \in \mathbb{R}^{(m-r)\times(m-r)}$ of $\widetilde{R}$ in (7). According to Corollary 2.1, for $m = O(T^{1/4-\varepsilon})$ with $\varepsilon \in (0, \frac{1}{4}]$, the estimate $\widetilde{\Pi}$ is a matrix of full-rank and also a consistent estimate of Π. Therefore the lower diagonal elements of $\widetilde{R}_{22}'$ are expected to be small. In particular, they converge to zero asymptotically at unit root speed $1/T$ as is shown in the following Theorem.

Theorem 2.2. *Let Assumptions 2.1–2.3 hold and $\widetilde{R}_1'$ denote the first r and by $\widetilde{R}_2'$ the last $m-r$ columns of $\widetilde{R}'$ in the QR-decomposition (7) of $\widetilde{\Pi}'$. Besides, define $\tilde{\mu}_k = \sqrt{\sum_{j=k}^m \widetilde{R}(k,j)^2}$. Then for $m = O(T^{1/4-\varepsilon})$ and $r = O(m^{\frac{1}{2\tau_1+1}})$ with $\varepsilon \in (0, \frac{1}{4}]$*

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

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

$$\tilde{\mu}_k = O_p(\frac{1}{T}) \quad k = r+1, \ldots, m$$

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

The first part of Theorem 2.2 provides identification of the cointegration space spanned by β. In the respective rate, however, unit root speed is generally slowed down by m and r^{τ_1} which is larger the faster $\sigma_r(\alpha)$ approaches zero in Assumption 2.3. But the subspace distance between $\widetilde{S}_1$ and β converges at a faster rate than the distance between $\widetilde{R}_1$ and α'. This is the key point in order to disentangle stationary and nonstationary components. Moreover, from point 2 of Theorem 2.2, the l_2-type weight $\tilde{\mu}_k$ achieves exact unit root speed for the irrelevant parts without affecting identification of the loadings α in speed of convergence. Therefore, $\tilde{\mu}_k$ yields a clearer separation of relevant and irrelevant columns and is the preferred weight for an adaptive Lasso procedure. Note that Theorem 2.2 contains the fixed dimensional case as a special case, where identification of the space of β from $\widetilde{S}_1$ is at unit root speed and the standard stationary speed $1/\sqrt{T}$ is obtained for the loadings.

The statistical properties of the QR-components of $\widetilde{\Pi}$ derived in Theorem 2.2 inspire the construction of the following adaptive group Lasso objective function (8) with group-wise weights for the determination of the cointegration rank (see Wei and Huang (2010) for group Lasso in the standard univariate i.i.d case). Hence columns $\hat{R}'(.,j)$ of the adaptive group-Lasso estimator $\hat{R}'$ minimize the following column-wise criterion over all $R'(.,j)$ for $j = 1, \ldots, m$

$$\sum_{t=1}^T \|\Delta\widetilde{Y}_t - R'\widetilde{S}'\widetilde{Y}_{t-1}\|_2^2 + \sum_{j=1}^m \frac{\lambda_T^{rank}}{\tilde{\mu}_j^\gamma}\|R'(.,j)\|_2 \tag{8}$$

where the penalization parameter λ_T^{rank} and the weight γ for adaptiveness in (8) 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 (8) as $\hat{r} = \text{rank}(\hat{R})$, where $\text{rank}(\hat{R})$ equals the number of non-zero columns in $\hat{R}'$.

This adaptive group Lasso procedure (8) exploits that according to Theorem 2.2 the last $m-r$ columns of $\widetilde{R}'$ converge to zero at a rate faster than the rate of the first r stationary columns for the stated conditions on m and r. With this, we can construct adaptive weights for a model selection consistent group Lasso procedure, which put a faster diverging penalty on any element in the space orthogonal to β and less on those stationary components in the cointegration space.

Remark 2.1. According to Theorem 2.2, the subspace distance between $\widetilde{S}_1$ and β converges at a faster rate than the subspace distance of $\widetilde{R}_1$ and α under the given conditions on m and r. Therefore the first step estimation error from using $\widetilde{S}$ in (8) instead of the infeasible true S_1 is negligible for estimating R from the Lasso criterion.

Moreover, even when m and r are both fixed, our approach features several advantages compared with existing literature: Firstly, the employed QR-decomposition is always real-valued without further constraints on the matrix $\widetilde{\Pi}$. Thus the Lasso criterion (8) only contains real-valued elements and can be minimized with standard optimization techniques. In comparison, a corresponding eigenvalue decomposition of an asymmetric matrix as e.g. in Liao and Phillips (2015) would in general contain complex values leading to a non-standard harmonic function optimization problem in a respective Lasso objective function. Secondly, after the QR-transformation based on the consistent pre-estimator, the objective function (8) has the same penalized representation as standard Lasso problem and is therefore straightforward to implement with any available numerically efficient algorithm. So our method is direct and ready to use.

The following theorem provides the statistical properties of adaptive group Lasso estimate from (8).

Theorem 2.3. *Under Assumptions 2.1–2.3 and if* λ_T^{rank} *satisfies* $\frac{\lambda_T^{rank}}{\sqrt{T}} r^{\tau_1\gamma+1/2} \to 0$ *and* $\frac{\lambda_T^{rank} T^{\gamma-1}}{m^{3/2}} \to \infty$, $m = O(T^{1/4-\varepsilon})$ *with* $\varepsilon \in (0, 1/4]$, *and* $r = O(m^{\frac{1}{2\tau_1+1}})$. *Then the solution* $\hat{R}$ *of the adaptive group Lasso criterion (8) satisfies*

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

Theorem 2.3 shows in part 1 rank selection consistency of the adaptive group Lasso technique for all admissible penalties λ_T^{rank} satisfying $\lambda_T^{rank} = o(\frac{\sqrt{T}}{r^{\tau_1\gamma+1/2}})$ and $\frac{m^{3/2}}{T^{\gamma-1}} = o(\lambda_T^{rank})$. Under our assumption on the explosion rate of m and r, setting e.g. γ as 2 allows for a large set of possible λ_T^{rank} choices even if the exact rate of r is unknown. Generally, the best finite sample performance is achieved if γ is not too large as also standard in the literature on stationary adaptive Lasso. Please see also our finite sample results in Section 5.

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

In part two of the Theorem, we get as a by-product to consistent cointegration rank selection also consistent estimates for α from the adaptive group Lasso criterion (8). Note that the obtained rate of convergence coincides with the infeasible oracle rate in the high-dimensional case when the true cointegration rate was known. In the case of fixed r and m we recover the standard stationary $T^{1/2}$-rate of convergence.

3. Lag selection

As for the rank choice, the standard VECM equation (2) is transformed in a Frisch–Waugh pre-step in order to focus on the lag selection. In particular, the effect of the nonstationary term Y_{-1} is discarded by employing $C = I_T - Y_{-1}'(Y_{-1}Y_{-1}')^{-1}Y_{-1}$ in (2)

$$\Delta\breve{Y}_t = B\Delta\breve{X}_{t-1} + \breve{w}_t \tag{9}$$

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

For lag selection, we obtain the least squares estimator $\breve{B}$ and the Ridge estimator $\widetilde{B}$ of the transient lag components B from (9) as

$$\breve{B} = \left(\frac{1}{T}\sum_{t=1}^T \Delta\breve{Y}_t\Delta\breve{X}_{t-1}'\right)\left(\frac{1}{T}\sum_{t=1}^T \Delta\breve{X}_{t-1}\Delta\breve{X}_{t-1}'\right)^{-1} \tag{10}$$

$$\widetilde{B} = \left(\frac{1}{T}\sum_{t=1}^T \Delta\breve{Y}_t\Delta\breve{X}_{t-1}'\right)\left(\frac{1}{T}\sum_{t=1}^T \Delta\breve{X}_{t-1}\Delta\breve{X}_{t-1}' + \frac{\lambda_T^{ridge}}{T} I_{mP}\right)^{-1}. \tag{11}$$

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

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

Theorem 3.1. *Let Assumptions 2.1–2.3 hold and $\breve{B}$ and $\widetilde{B}$ are as defined in (10) and (11). Assume $m = O(T^{\frac{1}{4}-\varepsilon})$ with $\varepsilon \in (0, 1/4]$. Then*

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

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

if $\lambda_T^{ridge} = o(\sqrt{T})$ for $\widetilde{B}$ in (11).

Note that all components in both estimators depend on the initial transformation C. Therefore for the consistency rates in Theorem 3.1 explicit rates of all blocks in (10) and (11) are crucial and therefore derived in the technical Lemma 3 in the online supplementary.

From Theorem 3.1 we obtain consistency results in the l_∞ norm for the vectorized coefficient matrices B_j with $j = 1, \ldots, P$ of the stationary transient components in (13). In contrast to the rank selection case, for the stationary lag coefficient pre-estimates there is no difference in speed between true zero coefficient matrices and non-zero ones only estimated as zero as in standard stationary adaptive Lasso selection problems. Thus we adopt the l_∞ norm in order to carefully ensure that if there exists at least one non-zero element in a coefficient matrix, the corresponding lagged term is relevant to the model. Compared to l_2 or Frobenius norm, l_∞ increases with the dimension m only at the logarithmic rate and is independent of the sparsity structure. It is therefore preferred as weight for the adaptive step.

Thus the adaptive Lasso estimate $\hat{B} = (\hat{B}_1, \ldots, \hat{B}_P)$ of the lag coefficient matrices in (9) minimizes the following objective function in lag coefficient matrices $B_j \in \mathbb{R}^{m\times m}$ of $B = (B_1, \ldots, B_P) \in \mathbb{R}^{m\times mP}$

$$\sum_{t=1}^{T} \|\Delta\breve{Y}_t - \sum_{j=1}^{P} B_j \Delta\breve{Y}_{t-j}\|_2^2 + \lambda_T^{lag} \sum_{j=1}^{P} \|vec(\widetilde{B}_j)\|_\infty^{-\gamma} \|B_j\|_F \tag{12}$$

As in the case of rank selection, a lag k should be included into the model, whenever $\hat{B}_k$ from the Lasso selection (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 $\hat{p} = \max_{1\leqslant k\leqslant P}\{k|\hat{B}_k \neq 0\}$. We also define the estimated active set $I_{\hat{B}}$ of (12) as the set of indices with non-zero $\hat{B}_j$ for $1 \leqslant j \leqslant p$, i.e., $I_{\hat{B}} = \{j|\hat{B}_j \neq 0\}$ and $\hat{B}_0 = (\hat{B}_j)_{j\in I_B} \in \mathbb{R}^{m\times lm}$ with $l \leqslant p$ consists of estimated coefficient matrices of the true active set I_B.

In the objective function (12), we penalize each coefficient matrix jointly by group Lasso rather than penalizing each element in the matrix separately. Such elementwise Lasso would be less robust in finite sample performance and potentially lead to problems in economic interpretation.

Remark 3.1. In the adaptive weight, theoretically also the use of the least-squares estimate $\breve{B}$ is justified yielding the same consistency result as below for the Ridge estimate $\widetilde{B}$. For a numerically stable adaptive Lasso procedure in finite samples, however, the use of the Ridge weight is essential in order to mitigate the large impact of multicollinearity effects. Also pre-estimates from an elastic net type procedure (see Zou and Hastie (2005)) or sure independence screening (see Fan and Lv (2008)) could be employed for a numerically stable weight in (12). Their detailed treatment, however, is beyond the scope of this paper.

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

Theorem 3.2. *Let Assumptions 2.1–2.3 hold. Moreover, $\frac{\lambda_T^{lag}}{\sqrt{T}} \to 0$ and $\frac{\lambda_T^{lag} T^{\frac{1}{2}(\gamma-1)}}{m^2(\log m)^{\gamma/2}} \to \infty$, $m = O(T^{1/4-\varepsilon})$, then for the solution $\hat{B}$ of (12) with I_B, $I_{\hat{B}}$ and B_0, $\hat{B}_0$ as defined below (9) it holds that*

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

Theorem 3.2 shows lag selection consistency together with consistency of the obtained adaptive Lasso estimates $\hat{B}_0$ for m diverging not too fast. This implies also consistency of the estimated lag length $\hat{p}$ from (12). Note that also nonconsecutive lags are identified.

Note that for model selection consistency in the lag there is no impact of the fact that the true rank r is unknown. Technically this is because after C transformation, the effect of the stationary component $Z_{1,t-1}$ is filtered out and the

non-stationary $Z_{2,t-1}$ decays to zero. Therefore, the rank just appears in the second order effect, see Lemma 3 in the online supplementary for details.

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

4. Rank selection for weakly dependent error terms

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

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

$$\Delta Y_t = \Pi Y_{t-1} + u_t \tag{13}$$

for $t = 1, \ldots, T$ where the dimension m of Y_t and rank r of $\Pi = \alpha\beta'$ are diverging with T as in (1). But now, we allow for a general weakly dependent form of the error term u_t in (13).

Assumption 4.1. The error term has the representation $u_t = \sum_{j=0}^{\infty} A_j w_{t-j}$ with $A_0 = I_m$ where for the components it holds that

1. w_t satisfies Assumption 2.1.
2. The coefficient matrices satisfy $\sum_{j=1}^{\infty} j\|A_j\|_F < \infty$.

In Assumption 4.1, the coefficient matrices of this infinite moving average process u_t must decay to zero fast enough so that u_t is a weakly dependent multiple time series and thus the partial sums can still be approximated by a Wiener process element-wise. In high dimensional case $\|A_j\|_F$ converges to zero at some rate actually imposes some sparse structure on the coefficient matrices. In particular, we get the following functional convergence result.

Theorem 4.1. *Let Assumption 4.1 hold. Then each element in u_t has bounded $(4+\delta)$th moment as the original innovation e_t. Besides, the partial sum of each u_t^k can be approximated by Brownian motion, i.e.,*

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

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

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

We can directly obtain the least-squares estimator $\widetilde{\Pi}$ of Π for the simple VECM (13) as

$$\widetilde{\Pi} = (\sum_{t=1}^{T} \Delta Y_t Y_{-1}')(\sum_{t=1}^{T} Y_{t-1}Y_{t-1}')^{-1} \tag{14}$$

which coincides with the estimate from (4) for the no lag case $p = 0$. We derive its statistical properties by using the Q-transformation from (5) to distinguish the r stationary Z_1 from the $m-r$ nonstationary components Z_2 in $Z = QY = (Z_1', Z_2')'$. Note that the Q-transformed problem (13) simplifies in the rank only case to

$$\begin{aligned} \Delta Z_{1,t} &= \beta'\alpha Z_{1,t-1} + v_{1,t} \\ \Delta Z_{2,t} &= v_{2,t} \end{aligned} \tag{15}$$

with $v_t = Qu_t = (v_{1,t}', v_{2,t}')'$ where $v_1 \in \mathbb{R}^r$ and $v_2 \in \mathbb{R}^{m-r}$. Note that here $\mathbf{E}(v_t Z_{1,t-1}')$ is non-zero, due to the possible dependence in u_t and thus in v_t according to Assumption 4.1. This causes an endogeneity bias such that left subspace

generated by $\widetilde{\Pi}$ in (14) does no longer approximate α but $\alpha_\star$ defined as

$$\begin{aligned}\alpha'_\star &= \alpha' + \Sigma_{z1}^{-1}\Gamma_{v1z1}^{1\prime}(\alpha'\beta)^{-1}\alpha' + \Sigma_{z1}^{-1}\Gamma_{v2z1}^{1\prime}(\beta'_\perp\alpha_\perp)^{-1}\beta'_\perp \\ &= \alpha' + \Sigma_{z1}^{-1}\Gamma_{uz1}^{1\prime} \end{aligned} \tag{16}$$

with $\Gamma_{uz1}^1 = \mathbf{E}(u_t Z'_{1,t-1})$ and $\Sigma_{z1} = \mathbf{E}(Z_{1,t-1}Z'_{1,t-1})$. We also set $\Gamma_{viz1}^1 = \mathbf{E}(v_{it}Z'_{1,t-1})$ with $i \in \{1,2\}$. Though, for $\alpha_\star$ defined in (16) Assumption 2.3 is not sufficient to ensure non-singularity of $\alpha'_\star\alpha_\star$. Singularity, however, would affect rank selection consistency of the Lasso procedure since the estimation error for the relevant r basis would be inflated by an exactly zero smallest singular value of $\alpha_\star$. We therefore require the condition in part 1 of Assumption 2.3 not only for α but also for the biased object $\alpha_\star$. This is needed even in fixed dimensional case where an $\alpha_\star$ without full row-rank would increase the estimation error for $\widetilde{S}_1$ in the QR-decomposition (7) from unit root speed $\frac{1}{T}$ in Theorem 2.2 to only $\frac{1}{\sqrt{T}}$ which makes it indistinguishable from the stationary parts. Therefore we require the following assumption

Assumption 4.2. Let part 1 of Assumption 2.3 hold. Moreover, the singular values of $\alpha_\star$ satisfy $0 < \sigma_r(\alpha_\star) \leqslant \cdots \leqslant \sigma_1(\alpha_\star) < \infty$. And there exist $K_2 > 0$ and $\tau_2 > 0$ such that $r^{\tau_2}\sigma_r(\alpha_\star) \geqslant K_2$.

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

Let $\mathbf{M}(s)$ denote the m-dimensional martingale process defined in Theorem 4.1, where $\mathbf{M}_1(s)$ marks the first r elements and $\mathbf{M}_2(s)$ the last $m-r$ components.

Theorem 4.2. *Let Assumptions 2.2, 4.1 and 4.2 hold. With $D_T = diag\,(I_r,\ TI_{m-r})$ and $\widetilde{\Pi}$ from (14) define*

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

Moreover, denote

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

where $\Xi = (\beta'\alpha)^{-1}\Big((\beta'\alpha + I_r)\Gamma_{v2z1}^{1,\prime} + \Sigma_{v1v2} + \Gamma_{12}^0 + \int_0^1 d\mathbf{M}_1(s)\mathbf{M}'_2(s)\Big)$ and $\mathbf{V}_{ij} = (\int_0^1 d\mathbf{M}_i(s)\mathbf{M}'_j(s) + \Gamma_{ij}^0)(\int_0^1 \mathbf{M}_j(s)\mathbf{M}'_j(s)ds)^{-1}$ for $i,j = 1,2$ with $\Gamma^0 = \sum_{k=1}^\infty \mathbf{E}(v_t v'_{t-k})$ and all other elements as defined below (16).

Then for $r = O(m^{\frac{1}{2\tau_1+1}})$ it holds that

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

There are two main differences between this result and the independent case in Theorem 2.1. First, there is a bias term $\Gamma_{vz1}^1 \neq 0$ due to the correlation between u_t and Z_{t-1}. Second, the rate of convergence for the unit root part is slightly smaller due to the larger exponent in the $\log T$-term. Though, the driving denominator is still $T^{1/4}$ as before. Moreover, the rate restriction on r coincides with the i.i.d case since the inverse of $\beta'\alpha$ in Ξ causes the l_2-norm of Ξ and thus of $\Psi_{\star,12}$, $\Psi_{\star,22}$ to increase at rate of r^{τ_1}.

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

Theorem 4.3. *Let Assumptions 2.2, 4.1 and 4.2 hold and $\widetilde{R}'_1$ denote the first r and by $\widetilde{R}'_2$ the last $m-r$ columns of $\widetilde{R}'$ in the QR-decomposition (7) of $\widetilde{\Pi}'$ in (14). With $\tilde{\mu}_k = \sqrt{\sum_{j=k}^m \widetilde{R}(k,j)^2}$ for $m = O(T^{1/4-\varepsilon})$ and $r = O(m^{\frac{1}{2\tau_2+1}})$ where $\varepsilon \in (0, \frac{1}{4}]$ it holds that*

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

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

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

Theorem 4.3 shows that identification of the cointegration space occurs at a slightly slower speed of convergence as in the i.i.d-case of Theorem 2.2. Weak dependence in the innovation also slows down the convergence of the Lasso adaptive weights in the true zero parts from unit root speed to $\frac{r^{\tau_1}}{T}$. Both points make it harder for adaptive Lasso (8) to disentangle true stationary and nonstationary components. Technically, the difference in convergence rates of Theorems 2.2 and 4.3 results from the fact that for $\Psi_\star$ with the additional bias $\Gamma^1_{viz\,1}$ the l_2 bounds for blocks in Ψ cannot be attained. Convergence in the third part can only be attained for $\alpha_\star$ instead of α but the rate is unaffected.

Therefore, the same logic for the design of group Lasso weights from the i.i.d case can still be employed. Thus, we can still use the adaptive group Lasso objective function (8) for rank selection with a pre-estimate $\widetilde{S}$ from a QR-decomposition of $\widetilde{\Pi}$ in (14). As before, it yields a columnwise estimate of $\hat{R}'$ from which we can determine the cointegration rank. The statistical properties of this procedure are provided in the following theorem.

Theorem 4.4. *Under Assumptions 2.2, 4.1 and 4.2, if* λ_T^{rank} *satisfies* $\frac{\lambda_T^{rank}}{\sqrt{T}} r^{\tau_2\gamma+1/2} \to 0$ *and* $\frac{\lambda_T^{rank} T^{\gamma-1}}{m^{3/2} r^{\tau_1(\gamma+1)}} \to \infty$, $m = O(T^{1/4-\varepsilon})$ *with* $\varepsilon \in (0, 1/4]$, *and* $r = O(m^{\frac{1}{2\tau_2+1}})$, *then the solution* $\hat{R}$ *of the adaptive group Lasso criterion (8) with pre-estimate* $\widetilde{S}$ *from a QR-decomposition of* $\widetilde{\Pi}$ *in (14) satisfies*

1. $\mathbb{P}\left(\sum_{j=1}^m \mathbb{I}_{\hat{R}'(j)\neq 0} = r\right) \geqslant 1 - \bar{C}_0(\frac{m^{3/2} r^{\tau_1(\gamma+1)}}{\lambda_T^{rank} T^{\gamma-1}})^2$ *for some* $\bar{C}_0 < \infty$
2. $\|\hat{R}'_1 - \alpha_\star H\|_F = O_p(\sqrt{\frac{mr}{T}})$

for some orthonormal matrix H.

Theorem 4.4 shows that given our assumptions, even if the innovations are weakly dependent, rank selection is still consistent. The estimate of the loading matrix, however, only consistently identifies $\alpha_\star$ as defined in (16) which generally differs from α.

5. Simulations

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

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

Exact model specifications of Π in (1) are constructed randomly by first generating two orthonormal matrices $U, V \in \mathbb{R}^{m\times r}$. Such orthonormal matrices can be obtained from QR-decomposition or singular value decomposition of a matrix with each element drawn from a standard normal distribution. Then we randomly draw elements for an $r \times r$ diagonal matrix Λ from univariate standard normal until $\Pi = U\Lambda V'$ first satisfies Assumption 2.2. As the main focus of this paper is rank selection in a cointegrated model, in all set-ups coefficient matrices B_j are set as diagonal with elements also drawn from a univariate standard normal. In this section, we set $P = 3$ to reduce computational time. Innovations w_t in (1) are drawn from the standard Normal or t-distribution with degrees of freedom $df \in \{8, 20, 200\}$ fulfilling the moment condition of Assumption 2.1. We study different degrees of cross-sectional dependence, with banded covariance matrices of the innovations of the form $\Sigma_w = (\rho^{|i-j|})_{ij}$ for $\rho = 0.0, 0.2, 0.4, 0.6$. We consider different combinations of these parameters for sample sizes $T = 400, 800, 1200, 1600$.

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

$$\min_\lambda \log|\hat{\Sigma}_w(\lambda)| + \frac{\log T}{T}\|vec(A(\lambda))\|_0 \tag{17}$$

where $A = \hat{R}(\lambda)$ in rank selection and $A = \hat{B}(\lambda)$ in lag selection, and $\hat{\Sigma}_w(\lambda)$ denotes the sample covariance matrix of the residuals for λ from (3) or (9).

In the following tables, each cell contains the percentages XX/YY of correct model selections by solving (8) and (12) for $b = 100$ repetitions of the respective model, where XX denotes the number of correct rank selections while YY is the number of correct lag length identifications. When the model has no transient terms, there exists only one number XX representing rank selection results.

Table 1 studies the performance of the adaptive group Lasso procedure for $m = 20$ dimensions with true rank $r = 5$ and lag $p = 1$ with $\rho = 0$ in the cross-correlation of the innovations. From top to bottom the difficulty of the selection problem increases with less existing moments in the innovation terms. This is also reflected in the reported results with excellent overall performance except in extreme cases where $T^{1/4}$ is smaller than 5, but the treated dimension is $m = 20$. Here, the conditions for Lasso selection consistency with $m = o(T^{1/4})$ are hard to justify. Though performance of the Lasso procedure is still quite good but affected by heavier tails in the innovations in particular in the lag selection case. For the same setup of Π

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

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

Table 2
Model selection results for model 1 with $m = 20$, rank $r = 5$, lag $p = 1$ and $\gamma = 3$.

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

Table 3
Model selection results for model 1 with $m = 20$, rank $r = 5$ and lag $p = 1$ for different ρ cross-section dependence with different γ-choices.

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

and B as in Table 1, we report model selection results for an almost normal type of innovation with $df = 200$ and substantial tail thickness $df = 20$ across different levels of strength in the cross-sectional correlation Σ_w in Table 2. The results show that even for substantial correlation with $\rho = 0.6$, performance is reliable for $T \geqslant 800$ even in the case of for $df = 20$ innovations with excess-kurtosis of 0.375. Generally, a larger degree of freedom leads to better rank selection results given the same T and ρ. Besides, simulations show that the size of ρ has a significant effect on model selection, which highlights the importance of Assumption 2.1 on the structure of Σ_w, i.e., the column-wise sums of absolute values must converge fast enough.

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

Besides, we present the estimation error of the loading matrix $\hat{R}_1$ and the cointegrating space $\tilde{S}_1$ in Fig. 1 for $df = 20$ and in Fig. 2 for $df = 200$ in the case $\rho = 0.0$. Because α and β are only unique up to rotation, the estimation error here is measured by using orthogonal projection matrices to uniquely identify subspace distances. In particular, we employ the R package LDRTools based on average orthogonal projection matrices proposed by Liski et al. (2016). The left bar in each plot corresponds to $T = 800$ and the right one to $T = 1200$. The estimation error for the cointegrating space is significantly smaller than that for the loading matrix due to the faster rate of convergence. Moving from sample size 800 to 1200 significantly improves results in both cases.

Model 2 uses the same Π as model 1 but considers only rank selection in VECM without transient dynamics, i.e. setting $B = 0$. Thus the problem is simpler and technically, the step of the Frisch–Waugh transformation by M in (3) can be omitted. The results can be found in Table 4. In small samples with $T = 400$ and for large ρ, this provides improvements in comparison to Table 2. Thus without lags, we get satisfactory performance even in these challenging cases of strong cross-sectional dependence.

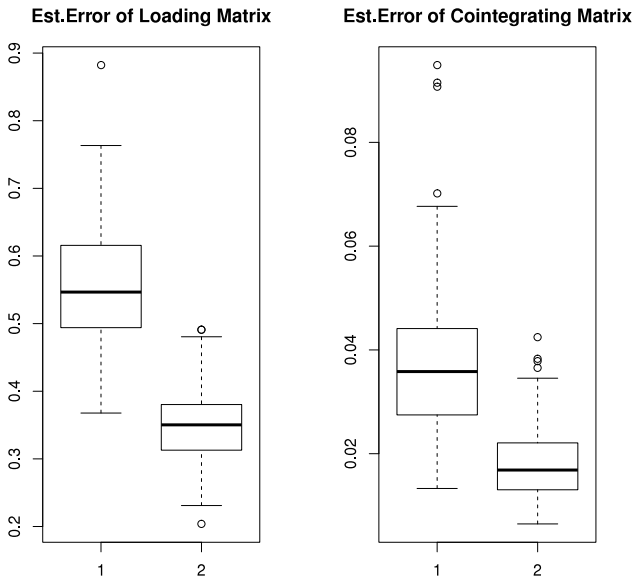

Fig. 1. Estimation Error of model 1 ($m = 20, r = 5, p = 1$) with t-distributed innovations and $df = 20$ for $\rho = 0$ setting $\gamma = 3$. Results are shown for $T = 800$ marked as case 1 on the x-axis and for case 2 of $T = 1200$.

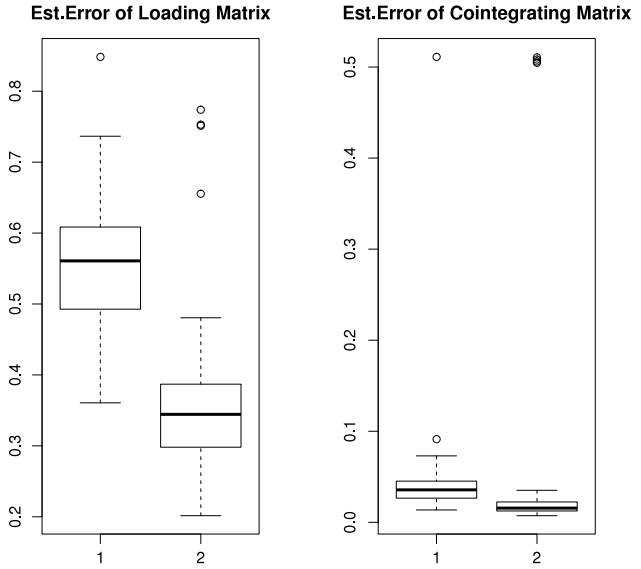

Fig. 2. Estimation Error of model 1 ($m = 20, r = 5, p = 1$) with t-distributed innovations and $df = 200$ for $\rho = 0$ setting $\gamma = 3$. Results are shown for $T = 800$ marked as case 1 on the x-axis and for case 2 of $T = 1200$.

Table 4
Model selection result for model 2 with $m = 20$, rank $r = 5$, lag $p = 0$ and $\gamma = 3$.

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

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

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

Table 6
Rank selection result for $m = 50$ with t-distributed innovations. $\rho = 0$ and $\gamma = 3$.

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

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

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

where w_t follows t-distribution with covariance $\Sigma_w = (\rho^{|i-j|})_{ij}$ as defined before. Besides, $A_1 = (a_{1,ij}) = (0.8^i \mathbb{I}_{i=j})$ and $A_2 = (a_{2,ij}) = ((-0.4)^i \mathbb{I}_{i=j})$ satisfy Assumption 4.1. As in Table 1, we set $\gamma = 3$ and choose λ by BIC. See Table 5 for results. When $T \geqslant 800$, the rank selection results are satisfactory, which is consistent with the theoretical results.

In Table 6, we present the rank selection results for the 50-dimensional case of model 3. Compare this to the usual simulation scenarios the high-dimensional non-stationary time series literature which usually do not go beyond dimension 20 (see e.g. Zhang et al. (2018)). We focus on results for innovations following a t-distribution with $df = 20$ and $df = 200$ respectively, with $\rho = 0.0$, i.e. $\Sigma_w = I_m$ only. For both cases, when $T \geqslant 2000$, the true rank can be estimated almost 100% correct. The increased sample size reflects the difficulty of the problem in dimensionality.

For the high-dimensional set-ups treated before, there exists no other valid feasible method for model determination against which we could evaluate our technique. Therefore, although our techniques are tailored to the high-dimensional case, we briefly illustrate that they can also be employed in standard low dimensions where benchmarks exist. In particular, we compare our methods with the Lasso-type techniques in Liao and Phillips (2015) using the "hardest" of their 2-dimensional models treated with $r = 1$ and $p = 3$. In particular, we set $\Pi = \begin{pmatrix} -1 & -0.5 \\ 1 & 0.5 \end{pmatrix}$ and $B_1 = B_3 = \text{diag}(0.4\ 0.4)$, $B_2 = 0$ and $\Sigma_w = \text{diag}(1.25\ 0.75)$. With 5000 simulation replications we get the following model selection results: for $T = 100$ we get 100%/86.14% while for $T = 400$ we obtain 100%/99.96% which compare to 99.54%/99.80% and 100%/99.98% by Table 2 in Liao and Phillips (2015). In their other settings, we also found similar comparable performance of the two techniques. Results are omitted here for the sake of brevity but are available on request.

6. Empirical example[4]

In this section, we employ our method to study the interconnectedness of the European sovereign and key players of the banking system during and after the financial crisis. We use CDS log prices of ten European countries and five selected financial institutions provided by Bloomberg terminal: *Germany, France, Belgium, Austria, Denmark, Ireland, Italy, Netherland, Spain, Portugal, BNP Paribas, SocGen Bank, LCL Bank, Danske Bank, Santander Bank*.[5] The sovereign countries we choose have different debt levels. The considered time span is from *Jan*. 1, 2013 to *Dec*. 31, 2016 with 1041 observations. *BNP Paribas, SocGen Banks* are chosen because they rank among the top three Europe based investment banks in Euro-Zone revenues. The other three banks are selected across EU countries covering the whole span from north to south and representing the variety of different financial market and general economic conditions. Initial Augmented Dicky Fuller tests show that the 15 variables are non-stationary but the first-order differences are stationary.

Fig. 1 suggests that there exists a strong co-movement among these components. Using our Lasso procedure, we find that there exist two cointegration relations. Fig. 2 gives an impression on the stable time evolution of these cointegrated series. Moreover, the time when the cointegrated series exhibit extreme values coincides with some important economic events.

4 All the figures for this section can be found in online supplementary.

5 UK is excluded due to Brexit.

Table 7
Each cell implies the contribution of variable denoted by its column name to the forecast error variance of the variable denoted by its row name. The number in row names is the horizon of the FEVD. The row denoted by *Sum* calculates the sum of each column except the element on the diagonal, which is the total contribution to all the other variables.

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

For example, in the middle of the year 2013, European countries were bargaining over the solution for the sovereign debt crisis while at the beginning of 2016 there occurred an economic slowdown in the key global economies.

To present the inter-connections among these 15-dimensional VECM components, we calculate the forecast error variance decomposition (FEVD hence after) due to the cointegrated part, i.e., the forecast error variance decomposition[6] derived from (3). From Table 7 reporting the FEVC results for a 5-and 10-step forecast horizon, we can conclude that leading economies in European Union, such as Germany, are neither risk-exporter nor risk-importer in the whole system.) Italy is the largest risk-exporter among the considered sovereign countries and Spain ranks second. Moreover, Italy and Spain have significant mutual influence on each other. The banks have stronger interconnectedness among themselves than with the sovereign countries. Moreover, Fig. 3 shows the contribution of Italy to the FEVD of other variables in the full horizon from step 0 to 30, which is consistent with the results in Table 7.

7. Conclusion

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

6 See e.g. Section 2.3.3 of Lütkepohl (2007).

Appendix A. Proofs

Proof of Theorem 2.1

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

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

and

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

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

We show (A.2) and (A.1) by studying blockwise elements of $\frac{1}{T}\sum_{t=1}^{T}\Delta\widetilde{Z}_t\widetilde{Z}'_{t-1}$ and $D_T^{-1}\frac{1}{T}\sum_{t=1}^{T}\widetilde{Z}_{t-1}\widetilde{Z}'_{t-1}$. Thus according to (6) we need to consider the following 8 different blocks.

1.+2. Purely stationary blocks $b_{11} = \frac{1}{T}\Delta Z_1 MZ'_{1,-1}$ and $\chi_{11} = \frac{1}{T}Z_{1,-1}MZ'_{1,-1}$

For the second block the standard law of large numbers argument from Lemma 2 yields:

$$\frac{1}{T}\sum_{t=1}^{T}\widetilde{Z}_{1,t-1}\widetilde{Z}'_{1,t-1} = \Sigma_{z1} - \Sigma_{z1\Delta x}\Sigma_{\Delta x}^{-1}\Sigma_{\Delta xz1} + R_1 \tag{A.3}$$

with $\|R_1\|_F = O_P(r/\sqrt{T})$. For the first term we get from (5)

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

This implies that

$$\|\frac{1}{T}\sum_{t=1}^{T}\Delta\widetilde{Z}_{1,t}\widetilde{Z}_{1,t-1} - (\beta'\alpha)\Sigma_{z1.\Delta x}\|_F = O_p(\frac{r}{\sqrt{T}}) \tag{A.4}$$

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

3. Mixed stationary/nonstationary block $b_{12} = \frac{1}{T}\Delta Z_1 MZ'_{2,-1}$

From (5) we get

$$\widetilde{Z}_{2,t} = \sum_{s=1}^{t}\widetilde{v}_{2,s} = \sum_{s=1}^{t}v_{2,s} - R_3 \tag{A.5}$$

with $\|R_3\|_F = O_P(r/\sqrt{T})$ since $\|\frac{1}{T}\sum_{s=1}^{T}v_{2,s}\Delta X'_{s-1}\|_2 = O_P(\frac{1}{\sqrt{T}})$ and Lemma 2. Thus $\frac{1}{T}\Delta Z_1 MZ'_{2,-1}$ can be further decomposed from (5) in ΔZ_1 by summation by part in $Z_{2,-1}$ as:

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

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

$$\|\frac{1}{T}\sum_{t=0}^{T}\Delta\widetilde{Z}_{1,t}\widetilde{Z}'_{2,t-1} + \Sigma_{v1v2}\|_F = O_p(\frac{\sqrt{mr}}{\sqrt{T}}) \tag{A.6}$$

4. Mixed stationary/nonstationary $b_{21} = \frac{1}{T}\Delta Z_2 M Z'_{1,-1}$

With (A.5) it holds that $\|\frac{1}{T}\sum_{t=1}^{T}\Delta\widetilde{Z}_{2,t}\widetilde{Z}'_{1,t-1}\|_F = \|\frac{1}{T}\sum_{t=1}^{T} v_{2,t}\widetilde{Z}'_{1,t-1}\|_F + O_p(\frac{\sqrt{mr}}{\sqrt{T}})$ which leads to

$$\|\frac{1}{T}\sum_{t=1}^{T}\Delta\widetilde{Z}_{2,t}\widetilde{Z}'_{1,t-1}\|_F = O_p(\frac{\sqrt{mr}}{\sqrt{T}}) \tag{A.7}$$

due to the independence condition in Assumption 2.1 and Lemma 2.

5. Purely nonstationary block $b_{22} = \frac{1}{T}\Delta Z_2 M Z'_{2,-1}$

From (A.5) we have

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\widetilde{Z}_{2,t-1}\Delta\widetilde{Z}_{2,t} = \frac{1}{T}\sum_{t=1}^{T}\widetilde{Z}_{2,t-1}v'_{2t} \\ &= \frac{1}{T}\sum_{t=1}^{T}Z_{2,t-1}v'_{2t} - (\frac{1}{T}\sum_{t=1}^{T}v_{2,t}\Delta X'_{t-1})(\frac{1}{T}\sum_{t=1}^{T}\Delta X_{t-1}\Delta X'_{t-1})^{-1}\frac{1}{T}\sum_{t=1}^{T}\Big((\sum_{s=0}^{t-1}\Delta X_s)v'_{2t}\Big)\end{aligned} \tag{A.8}$$

In this proof we focus on the term $\frac{1}{T}\sum_{t=1}^{T}Z_{2,t-1}v'_{2t}$ because the second term contains $\frac{1}{T}\sum_{t=1}^{T}v_{2,t}\Delta X'_{t-1}$ whose l_2 norm decaying to zero at the rate of $\sqrt{T}$, as proved in Lemma 2. The term $\frac{1}{T}\sum_{t=1}^{T}\Big((\sum_{s=0}^{t-1}\Delta X_s)v'_{2t}\Big)$ has similar performance as the first term on RHS of (A.8) whose property relies on the results in Theorem 4.1. Therefore, the LHS of (A.8) is dominated by the first term on the RHS.

For each $i, j = 1\ldots, m-r$ in the leading term on the right of (A.8),

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}Z^i_{2,t-1}v^j_{2,t} - \int_0^1 \mathbf{M}_{2,i}(s)d\mathbf{M}_{2,j}(s) \\ &= \sum_{t=1}^{T}\Big(\frac{1}{\sqrt{T}}Z^i_{2,t-1}(\frac{1}{\sqrt{T}}v^j_{2,t} - \int_{\frac{t-1}{T}}^{\frac{t}{T}}d\mathbf{M}_{2,j}(s)) + \int_{\frac{t-1}{T}}^{\frac{t}{T}}[\frac{1}{\sqrt{T}}Z^i_{2,t-1} - \mathbf{M}_{2,i}(s)]d\mathbf{M}_{2,j}(s)\Big) \\ &=_d \sum_{t=1}^{T}\Big(\frac{1}{\sqrt{T}}Z^i_{2,t-1}(\frac{1}{\sqrt{T}}v^j_{2,t} - \mathbf{M}_{2,j}(\frac{1}{T})) + \int_{\frac{t-1}{T}}^{\frac{t}{T}}[\frac{1}{\sqrt{T}}Z^i_{2,t-1} - \mathbf{M}_{2,i}(s)]d\mathbf{M}_{2,j}(s)\Big)\end{aligned} \tag{A.9}$$

To bound the first term on the RHS of (A.9), we apply integration by parts. Define $h^j_t = \frac{1}{\sqrt{T}}v^j_{2,t} - \mathbf{M}_{2,j}(\frac{1}{T})$ and $H^j_t = \sum_{s=1}^{t}h^j_t$, then

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

By strong invariance principle (see Theorem 12.7 of DasGupta (2008)), $sup_{t\leqslant T}|H^j_t| = O_{a.s.}(\frac{(\log T)^{1/2}(\log\log T)^{1/4}}{T^{1/4}})$, which provides an upper bound for variance of the middle term. Therefore, we can conclude that

$$|\sum_{t=1}^{T}\frac{1}{\sqrt{T}}Z^i_{2,t-1}h^j_t| = O_{a.s.}(\frac{(\log T)^{1/2}(\log\log T)^{1/4}}{T^{1/4}})$$

To bound the second term on the RHS of (A.9), we derive the upper bound for the integrand as

$$\begin{aligned}&\sup_{t\leqslant T,\,\frac{t-1}{T}\leqslant s\frac{t}{T}}|\frac{1}{\sqrt{T}}Z^i_{2,t-1} - \mathbf{M}_{2,i}(s)| \\ &\leqslant \sup_{t\leqslant T}|\frac{1}{\sqrt{T}}Z^i_{2,t-1} - \mathbf{M}_{2,i}(\frac{t-1}{T})| + \sup_{\frac{t-1}{T}\leqslant s\leqslant\frac{t}{T}}|\mathbf{M}_{2,i}(s) - \mathbf{M}_{2,i}(\frac{t-1}{T})| \\ &= O_{a.s.}(\frac{(\log T)^{1/2}(\log\log T)^{1/4}}{T^{1/4}} + \sqrt{\frac{\log T}{T}})\end{aligned} \tag{A.10}$$

where the first term on the RHS of (A.10) comes from strong invariance principle and second term from Levy modulus of continuity. Therefore by Ito isometry,

$$\begin{aligned}&\mathbf{E}\Big(\sum_{t=1}^{T}\int_{\frac{t-1}{T}}^{\frac{t}{T}}[\frac{1}{\sqrt{T}}Z^i_{2,t-1}-\mathbf{M}_{2,i}(s)]d\mathbf{M}_{2,j}(s)\Big)^2\\ =&\sum_{t=1}^{T}\mathbf{E}\Big(\int_{\frac{t-1}{T}}^{\frac{t}{T}}[\frac{1}{\sqrt{T}}Z^i_{2,t-1}-\mathbf{M}_{2,i}(s)]d\mathbf{M}_{2,j}(s)\Big)^2\\ =&\sum_{t=1}^{T}C_j\int_{\frac{t-1}{T}}^{\frac{t}{T}}\mathbf{E}[\frac{1}{\sqrt{T}}Z^i_{2,t-1}-\mathbf{M}_{2,i}(s)]^2ds=O\Big(\frac{(\log T)^{1/2}(\log\log T)^{1/4}}{T^{1/4}}\Big)^2\to 0\end{aligned}$$

where C_j is a constant depending on the $Var(\mathbf{M}_{2,j}(1))$.

Therefore, according to Proposition 1.26 on Page 131 of Revuz and Yor (1991), we have

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

All the convergence results are almost sure convergence, so the convergence result holds for each i, j uniformly (Lemma 9 of Zhang et al. (2018)), i.e.,

$$\sup_{i,j=1,2,\ldots,m-r}|\frac{1}{T}\sum_{t=1}^{T}\widetilde{Z}^i_{2,t-1}v^j_{2,t}-\int_0^1\mathbf{M}_{2,i}(s)d\mathbf{M}_{2,j}(s)|=O_{a.s.}(\frac{(\log T)^{1/2}(\log\log T)^{1/4}}{T^{1/4}})$$

because the union of countable non-convergence sets with measure zero is still a zero-measure set. Therefore,

$$\|\frac{1}{T}\sum_{t=1}^{T}v_{2,t}\widetilde{Z}'_{2,t-1}-\int_0^1 d\mathbf{M}_2(s)\mathbf{M}_2(s)'\|_F=O_p(\frac{m(\log T)^{1/2}(\log\log T)^{1/4}}{T^{1/4}}) \tag{A.11}$$

Thus from Eqs. (A.4)–(A.11) for the blocks $b_{11}, b_{12}, b_{21}, b_{22}$ we get the first part (A.1) of the initial claim.

6. Mixed stationary/nonstationary block $\chi_{12}=\frac{1}{T}Z_{1,-1}MZ'_{2,-1}$

From (5) we get with negligible R_7 that

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

Rearranging yields

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

As the first term on the right has been treated in block 3 above we can use (A.6). For the second term, the standard Brownian motion limit result applies. Moreover, we use that by Assumption 2.3 we have $\|(\beta'\alpha)^{-1}\|_2=O(r^{\tau_1})$. Hence in total, we find

$$\begin{aligned}&\|\frac{1}{T}\sum_{t=1}^{T}\widetilde{Z}_{1,t-1}\widetilde{Z}'_{2,t-1}+(\beta'\alpha)^{-1}(\Sigma_{v1v2}+\int_0^1 d\mathbf{M}_1\mathbf{M}'_2)\|_F\\ =&O\Big(r^{\tau_1}\sqrt{\frac{mr}{T}+\frac{mr(\log T)(\log\log T)^{1/2}}{\sqrt{T}}}\Big)\end{aligned} \tag{A.12}$$

7. Mixed stationary/nonstationary block $\chi_{21}=\frac{1}{T}\left(\frac{1}{T}Z_{2,-1}MZ'_{1,-1}\right)$

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

$$\|\frac{1}{T^2}\sum_{t=1}^{T}\widetilde{Z}_{2,t-1}Z'_{1,t-1}\|_F=O_p(\frac{\sqrt{mr}r^{\tau_1}}{T}) \tag{A.13}$$

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

Similar to b_{22} from block 5 we replace $\widetilde{Z}_{2,t-1}$ with $Z_{2,t-1}$ but with an additional T in the denominator from the pre-multiplication of D_T^{-1} in (A.2). Thus we get for any integer i, j ranging from 1 to $m - r$

$$\begin{aligned}&\frac{1}{T^2}\sum_{t=1}^{T} Z^i_{2,t-1}Z^j_{2,t-1} - \int_0^1 \mathbf{M}_{2,i}(s)\mathbf{M}_{2,j}(s)ds \qquad (A.14)\\ &= \sum_{t=1}^{T}\int_{\frac{t-1}{T}}^{\frac{t}{T}} \frac{1}{\sqrt{T}}Z^i_{2,t-1}\Big(\frac{1}{\sqrt{T}}Z^j_{2,t-1} - \mathbf{M}_{2,j}(s)\Big)ds + \sum_{t=1}^{T}\int_{\frac{t-1}{T}}^{\frac{t}{T}}\Big(\frac{1}{\sqrt{T}}Z^i_{2,t-1} - \mathbf{M}_{2,i}(s)\Big)\mathbf{M}_{2,j}(s)ds\end{aligned}$$

By the same argument as in block 5, we get

$$\sup_{i,j=1,2,\ldots,m-r} \Big|\frac{1}{T^2}\sum_{t=1}^{T} Z^i_{2,t-1}Z^j_{2,t-1} - \int_0^1 \mathbf{M}_{2,i}(s)\mathbf{M}_{2,j}(s)ds\Big| = O_{a.s.}\Big(\frac{(\log T)^{1/2}(\log\log T)^{1/4}}{T^{1/4}}\Big)$$

and therefore

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

Combining the blockwise results (A.3), (A.12)–(A.14) for χ_{11}, χ_{12}, χ_{21}, χ_{22} we get the second part of the initial claim (A.2).

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

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

Note that any term containing χ_{21} is of smaller order than the others as $\|\chi_{21}\|_F = O_p(\frac{\sqrt{mr}r^{\tau_1}}{T})$ due to (A.13). Therefore we find with (A.3), (A.12)–(A.14) and Lemma 1 that

$$\begin{aligned}&\|\xi_{11} - \Sigma^{-1}_{z1.\Delta x}\|_F = O_p\Big(\frac{r}{\sqrt{T}}\Big)\\ &\|\xi_{12} - \Sigma^{-1}_{z1.\Delta x}(\beta'\alpha)^{-1}\big(\Sigma_{v1v2}(\int_0^1 \mathbf{M}_2(s)\mathbf{M}'_2(s)ds)^{-1} + \mathbf{V}_{12}\big)\|_F = O_p\Big(r^{\tau_1}\sqrt{\frac{mr}{T} + \frac{mr(\log T)(\log\log T)^{1/2}}{\sqrt{T}}}\Big)\\ &\|\xi_{21}\|_F = O_p\Big(\frac{\sqrt{mr}r^{\tau_1}}{T}\Big)\\ &\|\xi_{22} - (\int_0^1 \mathbf{M}_2(s)\mathbf{M}'_2(s)ds)^{-1}\|_F = O_p\Big(m\frac{(\log T)^{1/2}(\log\log T)^{1/4}}{T^{1/4}}\Big) \qquad (A.15)\end{aligned}$$

Thus we get for $\widetilde{\Psi} = (\frac{1}{T}\sum_{t=1}^{T}\Delta\widetilde{Z}_t\widetilde{Z}'_{t-1})\xi$ from (A.1) and (A.15) together with Lemma 1 and the assumption $r = O(m^{\frac{1}{2\tau_1+1}})$ that

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

Proof of Corollary 2.1

Proof. The proof follows directly from Theorem 2.1 with $\Psi_0 = E(\Psi)$ and the weak law of large numbers. $\square$

Proof of Theorem 2.2

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

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

We pre-multiply $\widetilde{\Pi}'$ by matrix β_0. Thus we get with $\widetilde{\Psi} = Q\widetilde{\Pi}Q^{-1}D_T$ as in Theorem 2.1

$$\begin{aligned} \beta_0\widetilde{\Pi}' &= \begin{pmatrix} \beta'\widetilde{\Pi}' \\ \beta'_\perp\widetilde{\Pi}' \end{pmatrix} = \begin{pmatrix} I_r & \frac{1}{T}\beta'\alpha_\perp \\ 0 & \frac{1}{T}\beta'_\perp\alpha_\perp \end{pmatrix} \left(Q^{-1}\widetilde{\Psi}\right)' \\ &= \begin{pmatrix} I_r & \frac{1}{T}\beta'\alpha_\perp \\ 0 & \frac{1}{T}\beta'_\perp\alpha_\perp \end{pmatrix} \begin{pmatrix} \alpha(\beta'\alpha)^{-1}\widetilde{\Psi}_{11} + \beta_\perp(\alpha'_\perp\beta_\perp)^{-1}\widetilde{\Psi}_{21} & \alpha(\beta'\alpha)^{-1}\widetilde{\Psi}_{12} + \beta_\perp(\alpha'_\perp\beta_\perp)^{-1}\widetilde{\Psi}_{22} \end{pmatrix}' \end{aligned} \tag{A.16}$$

For the left block of $\left(Q^{-1}\widetilde{\Psi}\right)'$ we use that by Theorem 2.1, $\|\widetilde{\Psi}_{21}\|_F = O_P\left(\sqrt{mr/T}\right)$ and that $\|\widetilde{\Psi}_{11} - (\beta'\alpha)\|_F = O_P\left(rT^{-1/2}\right)$. Therefore we get from this part for the first block on the left hand side of (A.16) that

$$\|\beta'\widetilde{\Pi}' - \alpha'\|_F = O_p\left(\frac{r}{\sqrt{T}} + \sqrt{\frac{mr}{T}}\right) = O_P\left(\sqrt{\frac{mr}{T}}\right) \tag{A.17}$$

For the second block on the left hand side of (A.16), note that $\widetilde{\Psi}_{12}$, $\widetilde{\Psi}_{22}$ have their l_2 norms diverging at the rate of $\sqrt{m}$ due to the stochastic integral part by random matrix theory (see Vershynin (2012)). Therefore we get

$$\|\beta'_\perp\widetilde{\Pi}'\|_2 = O_p(\frac{\sqrt{m}}{T})\,. \tag{A.18}$$

We now use (A.17) and (A.18) in order to prove the stated claims of the theorem in reverse order and start with part 2. Due to the unitary invariance property of singular values, we have

$$\sigma_j(\beta_0\widetilde{\Pi}') = \sigma_j(S\widetilde{\Pi}') = \sigma_j(\widetilde{R}) \tag{A.19}$$

for all $j = 1, \ldots, m$. With Eq. (A.17), this implies in particular that

$$|\sigma_j(\widetilde{R}) - \sigma_j(\alpha)| = O_p(\sqrt{\frac{mr}{T}}) \quad \text{for } j = 1, \ldots, r \tag{A.20}$$

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

The column-pivoting step in the QR decomposition makes the $\widetilde{R}_{11}$ a well-conditioned matrix, thus the largest r singular values in $\widetilde{R}$ are in $\widetilde{R}_1$ which contains the first r-rows. Besides, the strict upper-triangular structure of $\widetilde{R}$ excludes linear dependence between any two rows in $\widetilde{R}_1$. Therefore, we can conclude that

$$\sigma_r(\widetilde{R}) \leqslant \sqrt{\sum_{j=k}^{m} \widetilde{R}(k,j)^2} \leqslant \sigma_1(\widetilde{R}) \quad \text{for } k = 1, \ldots, r \tag{A.21}$$

The matrix perturbation theory result (A.20) provides further bounds for l_2 norm of each row in $\widetilde{R}_1$, i.e.,

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

In the same way we obtain from (A.19) together with (A.18), that

$$|\sigma_j(\widetilde{R})| = O_p(1/T) \tag{A.22}$$

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

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

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

and thus

$$|\sigma_j(\widetilde{R}_1) - \sigma_j(\alpha)| = O_p(\sqrt{\frac{mr}{T}}) \quad \text{for } j = 1, \ldots, r \tag{A.23}$$

Thus we have shown claim 3 of the theorem.

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

$$\begin{pmatrix} \beta'\widetilde{\Pi}' \\ \beta'_\perp\widetilde{\Pi}' \end{pmatrix} = \begin{pmatrix} \beta'\widetilde{S}_1\widetilde{R}_{11} & \beta'\widetilde{S}_1\widetilde{R}_{12} + \beta'\widetilde{S}_2\widetilde{R}_{22} \\ \beta'_\perp\widetilde{S}_1\widetilde{R}_{11} & \beta'_\perp\widetilde{S}_1\widetilde{R}_{12} + \beta'_\perp\widetilde{S}_2\widetilde{R}_{22} \end{pmatrix}. \tag{A.24}$$

By equating (A.16) and (A.24) we get

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

which is equivalent to

$$\begin{aligned} \beta'_\perp\widetilde{S}_1 = &-\begin{bmatrix} 0 & \beta'_\perp\widetilde{S}_2\widetilde{R}_{22} \end{bmatrix}\widetilde{R}'_1(\widetilde{R}_1\widetilde{R}'_1)^{-1} \\ &+ \frac{1}{T}(\beta'_\perp\alpha_\perp)\Big(\alpha(\beta'\alpha)^{-1}\widetilde{\Psi}_{12} + \beta_\perp(\alpha'_\perp\beta_\perp)^{-1}\widetilde{\Psi}_{22}\Big)'\widetilde{R}'_1(\widetilde{R}_1\widetilde{R}'_1)^{-1} \end{aligned} \tag{A.25}$$

Note that for the first term on the right hand side of (A.25) we get due to (A.18) that $\|\beta'_\perp\widetilde{S}_2\widetilde{R}_{22}\|_2 = O_P(1/T)$. Therefore, the upper-bound in l_2 norm for $\beta'_\perp\widetilde{S}_1$ is driven by the rate of $(\widetilde{R}_1\widetilde{R}'_1)^{-1}/T$. From (A.23) we have that the singular values of $\widetilde{R}_1$ can be approximated by those of α. Therefore using Assumption 2.3 we conclude in total that $\|\beta'_\perp\widetilde{S}_1\|_F = O_p(\frac{mr^{2\tau_1}}{T})$. □

Proof of Theorem 2.3

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

Denote by $\widetilde{S}'\widetilde{Y}_{t-1} = \begin{bmatrix} \check{Z}_{1,t-1} \\ \check{Z}_{2,t-1} \end{bmatrix}$ where $\check{Z}_{1,t-1}$ is the projection of Y_{t-1} onto the subspace generated by $\widetilde{S}_1$. According to Theorem 2.2, the subspace distance between $\widetilde{S}_1$ and β converges at a faster rate ($\|\beta'_\perp\widetilde{S}_1\|_F = O_p(\frac{mr^{2\tau_1}}{T})$) than the subspace distance of $\widetilde{R}_1$ and α ($\sqrt{\frac{mr}{T}}$) under the given conditions on m and r. Therefore the first step estimation error from using $\widetilde{S}$ in (8) instead of the infeasible true S_1 is negligible and wlog. We use $\widetilde{Z}_{1,t-1}$ instead of $\check{Z}_{1,t-1}$ and $\widetilde{Z}_{2,t-1}$ instead of $\check{Z}_{2,t-1}$ (both are unit root process) for the rest of this proof for ease of notation. The replace of β by $\widetilde{S}_1$ is a common approach in cointegration literature, e.g. Ahn and Reinsel (1990) and Lütkepohl (2007) (Remark 3 on Page 293). Because $\widetilde{S}_1$ is a consistent estimator for the subspace generated by β, all the estimators for $Z_{2,t-1}$ are dominated by the unit root process only.

Since α and β are only identified up to rotation, we write wlog $\bar{\alpha} = \alpha H$ with H as defined for $\widetilde{S}_1$. Note that $\bar{\alpha}$ and α describe the same space. Define $\bar{\alpha}_0 = [\bar{\alpha}, 0_{m\times m-r}]$, $\delta_R = \hat{R}' - \bar{\alpha}_0$ and δ_{R_1} for the first r columns in δ_R. Then we have

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

Therefore, the minimization of (8) in $\hat{R}$ is equivalent to minimizing

$$\sum_{t=1}^{T} -2w'_t(\widetilde{Z}'_{t-1}\otimes I_m)vec(\delta_R) + vec(\delta_R)'(\widetilde{Z}_{t-1}\widetilde{Z}'_{t-1}\otimes I_m)vec(\delta_R) + \sum_{j=1}^{m}\frac{\lambda_T^{rank}}{\tilde{\mu}_j^{\gamma}}\|\bar{\alpha}_0(,j) + \delta_R(,j)\|_2 \tag{A.26}$$

in δ_R. With $D_{1T} = diag\{\sqrt{T}I_r, TI_{m-r}\}$ the term inside the first sum can be written as

$-2w_t'(\widetilde{Z}_{t-1}'D_{1T}^{-1}\otimes I_m)vec(\delta_R D_{1T}) + vec(\delta_R D_{1T})'(D_{1T}^{-1}\widetilde{Z}_{t-1}\widetilde{Z}_{t-1}'D_{1T}^{-1}\otimes I_m)vec(\delta_R D_{1T})$. Thus the Karush–Kuhn–Tucker (KKT) condition for group-wise variable selection from (A.26) is

$$-\frac{1}{\sqrt{T}}\sum_{t=1}^{T} w_t\widetilde{Z}_{1,t-1}' + \sqrt{T}\delta_{R1}\frac{1}{T}\sum_{t=1}^{T}\widetilde{Z}_{1,t-1}\widetilde{Z}_{1,t-1}' = -[\frac{\bar{\lambda}_{1,T}}{2\sqrt{T}}\frac{\bar{\alpha}(,1)}{\|\bar{\alpha}(,1)\|_2},\ldots,\frac{\bar{\lambda}_{r,T}}{2\sqrt{T}}\frac{\bar{\alpha}(,r)}{\|\bar{\alpha}(,r)\|_2}] \tag{A.27}$$

$$\|\Big(\sum_{t=1}^{T} -2\frac{1}{T}w_t\widetilde{Z}_{2,t-1}' + 2\sqrt{T}\delta_{R1}\frac{1}{T^{3/2}}\widetilde{Z}_{1,t-1}\widetilde{Z}_{2,t-1}'\Big)_j\|_2 < \frac{\bar{\lambda}_{r+j,T}}{T} \tag{A.28}$$

where $\bar{\lambda}_{j,T} = \frac{\lambda_T^{rank}}{\tilde{\mu}_j^{\gamma}}$ and the subscript j denotes the jth column. The expression follows since the derivative of the first term in (A.26) w.r.t. $vec(\delta_R D_{1T})$ is $\sum_{t=1}^{T} -2(D_{1T}^{-1}\widetilde{Z}_{t-1}\otimes I_m)w_t + 2(D_{1T}^{-1}\widetilde{Z}_{t-1}\widetilde{Z}_{t-1}'D_{1T}^{-1}\otimes I_m)vec(\delta_R D_{1T}) = \sum_{t=1}^{T} -2w_t\widetilde{Z}_{t-1}'D_{1T}^{-1} + 2\delta_R D_{1T}D_{1T}^{-1}\widetilde{Z}_{t-1}\widetilde{Z}_{t-1}'D_{1T}^{-1}$.

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

$$S_{z1z1} = \frac{1}{T}\sum_{t=1}^{T}\widetilde{Z}_{1,t-1}\widetilde{Z}_{1,t-1}' \qquad S_{wz1} = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} w_t\widetilde{Z}_{1,t-1}'$$

$$S_{z1z2} = \frac{1}{T^{3/2}}\sum_{t=1}^{T}\widetilde{Z}_{1,t-1}\widetilde{Z}_{2,t-1}' \qquad S_{wz2} = \frac{1}{T}\sum_{t=1}^{T} w_t\widetilde{Z}_{2,t-1}'$$

From the proof of Theorem 2.1, we can use that $S_{z1z2} = \chi_{12}/\sqrt{T}$ in block 6. Thus we can conclude from (A.12) that $\|S_{z1z2}\|_2 = O_p(\frac{r^{\tau_1}}{\sqrt{T}})$. Moreover, (A.3) and Lemma 2 imply that S_{z1z1} and S_{wz1} have bounded l_2 norm. Therefore we can re-write the first KKT-conditions (A.27) for the stationary part as

$$\sqrt{T}\delta_{R1} = -\frac{\lambda_T^{rank}}{2\sqrt{T}}V_\alpha S_{z1z1}^{-1} + S_{wz1}S_{z1z1}^{-1} \tag{A.29}$$

which implies that

$$\|\sqrt{T}\delta_{R1}\|_2 = O_p(\frac{\lambda_T^{rank}}{\sqrt{T}}r^{\tau_1\gamma+\frac{1}{2}} + 1) = o_P(1) \tag{A.30}$$

The convergence in (A.30) follows from the condition on the tuning parameter $\frac{\lambda_T^{rank}}{\sqrt{T}}r^{\tau_1\gamma+\frac{1}{2}} \to 0$ in the theorem. It thus yields that each element in δ_{R1} converges to zero at the rate of $\sqrt{T}$. Hence, the first r columns of the solution $\hat{R}'$ in (8) are $\sqrt{T}$-consistent for $\bar{\alpha}$.

Moreover, for the second part (A.28) of the KKT conditions, we plug in (A.29). Hence for the exclusion of the non-stationary components from (8), it is sufficient if

$$\|(S_{wz1}S_{z1z1}^{-1}S_{z1z2} - S_{wz2})_k\|_2 < \frac{\lambda_T^{rank}}{2T}\tilde{\mu}_{r+k}^{-\gamma} - \frac{\lambda_T^{rank}}{2\sqrt{T}}\|(V_\alpha S_{z1z1}^{-1}S_{z1z2})_k\|_2 \tag{A.31}$$

for $k = 1, 2, \ldots, m-r$. It remains to show that (A.31) is bounded in probability which implies selection consistency holds with probability one.

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

Thus the RHS of (A.31) is dominated by the first term. The LHS of (A.31) is dominated by S_{wz2} since $\|S_{wz1}S_{z1z1}^{-1}S_{z1z2}\|_2 = O_P(r^{\tau_1}/\sqrt{T})$ due to (A.3), (A.12), (A.6) and Lemma 2.

Moreover, for S_{wz2} we use that $\bar{Z}_{2,t} = \sum_{s=1}^{t}\alpha_\perp' w_s$ as in (A.5) to get for all $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, m-r$

$$\begin{aligned}\mathbf{E}(\frac{1}{T}\sum_{t=1}^{T} w_t^i\widetilde{Z}_{2,t-1}^j)^2 &= \frac{1}{T^2}\mathbf{E}(\sum_{t=1}^{T}(w_t^i)^2(\bar{Z}_{2,t-1}^j)^2) + \frac{1}{T^2}\mathbf{E}(\sum_{s\neq t}^{T} w_s^i\bar{Z}_{2,s-1}^j w_t^i\bar{Z}_{2,t-1}^j) + o_p(1) \\ &= \frac{1}{T^2}\sum_{t=1}^{T}\mathbf{E}((w_t^i)^2)\mathbf{E}((\bar{Z}_{2,t-1}^j)^2) + o_p(1) = O_p(\frac{1}{T^2}\sum_{t=1}^{T} t) = O_P(1)\end{aligned} \tag{A.32}$$

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

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

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

for some $0 < C_0 < \infty$ where we use Chebyshev's inequality and (A.32) together with $\widetilde{\mu}_{r+k} = O_P(1/T)$ from Theorem 2.2 in the last line. Thus with $\frac{m^{3/2}}{\lambda_T^{rank}T^{\gamma-1}} \to 0$ we simultaneously exclude the last $m-r$ columns with probability tending to 1. □

Proof of Theorem 3.1

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

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

Hence we can write the first component with $S^{z,-1}$ from (3) explicitly as

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

Thus Lemma 3 implies that

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

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

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

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

$$\|vec(\breve{B} - B) - vec\Big(\frac{1}{T}\sum_{t=1}^T w_t \Delta\dot{X}_{t-1}'\Big)\|_\infty \leqslant \|(\breve{B} - B) - \frac{1}{T}\sum_{t=1}^T w_t \Delta\dot{X}_{t-1}'\|_F = O_p\Big(\frac{m}{T^{3/2}}\Big). \tag{A.33}$$

According to the maximal inequality in Chernozhukov et al. (2013), we get

$$\Big\|\frac{1}{T}\sum_{t=1}^T vec(w_t \Delta\dot{X}_{t-1}')\Big\|_\infty = O_p\Big(\sqrt{\frac{\log m}{T}}\Big), \tag{A.34}$$

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

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

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

Proof of Theorem 3.2

Because the proof for lag selection consistency is similar to rank selection, we leave it in the online supplementary.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2018.09.018.

References

Ahn, S.K., Reinsel, G.C., 1990. Estimation for partially nonstationary multivariate autoregressive models. J. Amer. Statist. Assoc. 85 (411), 813–823.
Basu, S., Michailidis, G., 2015. Regularized estimation in sparse high-dimensional time series models. Ann. Statist. 43 (4), 1535–1567.
Bickel, P.J., Levina, E., 2008. Covariance regularization by thresholding. Ann. Statist. 2577–2604.
Bickel, P.J., Ritov, Y., Tsybakov, A.B., 2009. Simultaneous analysis of Lasso and Dantzig selector. Ann. Statist. 37 (4), 1705–1732.
Boswijk, H.P., Jansson, M., 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.
Cavaliere, G., Rahbek, A., Taylor, A.M.R., 2012. Bootstrap determination of the co-integration rank in vector autoregressive models. Econometrica 80 (4), 1721–1740.
Chao, J.C., Phillips, P.C., 1999. Model selection in partially nonstationary vector autoregressive processes with reduced rank structure. J. Econometrics 91 (2), 227–271.
Chernozhukov, V., Chetverikov, D., Kato, K., 2013. Comparison and anti-concentration bounds for maxima of gaussian random vectors. arXiv:1301.4807.
DasGupta, A., 2008. Asymptotic Theory of Statistics and Probability. In: Springer Texts in Statistics, Springer New York.
Engle, R., Granger, C., 1987. Co-integration and error correction: representation, estimation and testing. Econometrica 55, 257–276.
Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (5), 849–911.
Hubrich, K., Lütkepohl, H., Saikkonen, P., 2001. A review of systems cointegration tests. Econometric Rev. 20 (3), 247–318.
Johansen, S., 1988. Statistical analysis of cointegration vectors. J. Econom. Dynam. Control 12 (2–3), 231–254.
Johansen, S., 1991. Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. Econometrica 59 (6), 1551–1580.
Knight, K., Fu, W., 2000. Asymptotics for lasso-type estimators. Ann. Statist. 28 (5), 1356–1378.
Kock, A.B., Callot, L., 2015. Oracle inequalities for high dimensional vector autoregressions. J. Econometrics 186 (2), 325–344.
Kosorok, M.R., Ma, S., 2007. Marginal asymptotics for the large p, small n paradigm: With applications to microarray data. Ann. Statist. 35 (4), 1456–1486.
Li, H., Li, Q., Shi, Y., 2017. Determining the number of factors when the number of factors can increase with sample size. J. Econometrics 197 (1), 76–86.
Liao, Z., Phillips, P.C., 2015. Automated estimation of vector error correction models. Econom. Theory 31 (03), 581–646.
Liski, E., Nordhausen, K., Oja, H., Ruiz-Gazen, A., 2016. Combining linear dimension reduction subspaces. In: Proceedings of ICORS 2015.
Lütkepohl, H., 2007. New Introduction to Multiple Time Series Analysis. Springer Publishing Company, Incorporated.
Onatski, A., Wang, C., 2018. Alternative asymptotics for cointegration tests in large VARs. Econometrica 86 (4), 1465–1478.
Phillips, P.C., 2014. Optimal estimation of cointegrated systems with irrelevant instruments. J. Econometrics 178 (Part 2), 210–224, Recent Advances in Time Series Econometrics.
Revuz, D., Yor, M., 1991. Continuous Martingales and Brownian Motion. In: Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 293, Springer-Verlag, Berlin, p. x+533.
Signoretto, M., Suykens, J., 2012. Convex estimation of cointegrated VAR models by a nuclear norm penalty. IFAC Proc. 45 (16), 95–100.
Stewart, G.W., 1984. Rank degeneracy. SIAM J. Sci. Stat. Comput. 5 (2), 403–413.
Stewart, G.W., Sun, J., 1990. Matrix Perturbation Theory. Academic Press.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.
Vershynin, R., 2012. Introduction to the non-asymptotic analysis of random matrices. In: Eldar, Y.C., Kutyniok, G. (Eds.), Compressed Sensing: Theory and Applications. Cambridge University Press, pp. 210–268.
Wei, F., Huang, J., 2010. Consistent group selection in high-dimensional linear regression. Bernoulli 16 (4), 1369–1384.
Wilms, I., Croux, C., 2016. Forecasting Using Sparse cointegration. Int. J. Forecast. 32, 1256–1267.
Xiao, Z., Phillips, P.C., 1999. EFFICIENT detrending in cointegrating regression. Econom. Theory 15, 519–548.
Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat. Methodol. 68 (1), 49–67.
Zhang, R., Robinson, P., Yao, Q., 2018. Identifying cointegration by eigenanalysis. J. Amer. Statist. Assoc. 1–12.
Zhao, P., Yu, B., 2006. On model selection consistency of lasso. J. Mach. Learn. Res. 7 (Nov), 2541–2563.
Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101 (476), 1418–1429.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2), 301–320.