Universität St.Gallen

An Indirect Method For Cointegration Estimation

Author:
Jon Kqiku

Supervisor:
Prof. Dr. Matthias Fengler

Co-Supervisor:
Prof. Dr. Juan-Pablo Ortega

A thesis presented for the degree of
Master in Quantitative Economics and Finance

June 7, 2021

An Indirect Method For Cointegration Estimation

Jon Kqiku

Abstract

This thesis introduces a new method for estimating cointegrating relationships in multivariate stochastic processes. The novelty of this method comes from the fact that the cointegrating matrix is estimated in an indirect manner. Indeed, the method exploits Johansen's benchmark representation of a cointegrated system, the reduced VECM, to infer the cointegrating matrix from a VAR estimation using a matrix rank factorization, rather than estimating it directly using the likelihood of a reduced VECM representation, where the matrix of interest appears explicitly. Because of this indirect approach, the indirect cointegration estimator (ICE) is a generic and more flexible method in which any VAR estimator can be applied before retrieving the cointegrating matrix. As a consequence, and unlike Johansen's benchmark method, ICE is also applicable in high-dimensional and sparse settings by accordingly choosing a regularized estimator in the first step of VAR estimation. To assess the validity of this method, we perform a simulation study in which benchmark direct cointegration estimators (DiCE) and ICE using regularized VAR estimators are applied on simulated cointegrated processes. We find a similar performance between DiCE and ICE, validating the use of this method as an alternative way of estimating cointegration. The simulation study also highlights the difficulty of accurate cointegration estimation in fairly large systems even in cases where the system is known to be cointegrated.

Contents

Nomenclature

$\|\cdot\|_\infty$	L^∞-norm, $\|x\|_\infty = \max\{\|x_1\|, \ldots, \|x_n\|\}$
$\|\cdot\|_{\max}$	Max norm, $\|A\|_{\max} = \max_{i,j} \|a_{ij}\|$
$\|\cdot\|_p$	L^p-norm, $\|x\|_p = \left(\sum_{i=1}^n \|x_i\|^p\right)^{1/p}$
$\|\cdot\|_{\mathrm{F}}$	Frobenius norm, $\|A\|_{\mathrm{F}} = \|A\|_{2,2}$
$\|\cdot\|_{p,q}$	$L^{p,q}$-norm, $\|A\|_{p,q} = (\sum_{j=1}^n (\sum_{i=1}^m \|a_{ij}\|^p)^{q/p})^{1/q}$
$\otimes$	Kronecker product
$\top$	Transpose operator of a matrix
tr	Trace operator of a matrix
vec	Vectorization operator
$\mathrm{vec}_{n\times m}$	Memory vectorization, $\mathrm{vec}_{n\times m}^{-1}(\mathrm{vec}_{n\times m}(A)) = A$

1 Introduction

Cointegration is the property of an integrated multivariate stochastic process that there exists at least one linear combination of the variables in the system that yields a univariate process of inferior integration order than that of the system. In particular, if a system is $I(1)$, it is cointegrated if there exists a full rank matrix such that the linear transformation of the $I(1)$ vector process by that matrix is an $I(0)$, i.e. stationary, vector process. The interest for cointegration is sparked by its interpretation as a long-term equilibrium relationship between variables, which is of interest in fields such as economics, finance, meteorology, neuroscience and others. The first notable work on cointegration was done by Engle and Granger in [5] where they proposed a two-step method for estimating and testing for cointegration. The estimation is done by means of linear regression of one of the variables on all the others. In a previous work by Granger, this kind of regression where the variables exhibit unit roots and no true relationship was called spurious, since it often leads to high coefficients of determination and high significance on the coefficients solely because of the unit roots. In the above cited work, the regression is used as a first step to estimate potential cointegration coefficients, that are or not confirmed in a second step with a unit root test on the residuals of the regression. Indeed, the residual series can be seen as the cointegrating process if it is stationary, i.e. the unit root test is rejected. The most notable work on cointegration and still influential today came in a series of work by Johansen in which he shows that a system is cointegrated if and only if a matrix in the VECM representation of the system has not full rank. Its rank is called the cointegration rank and equals the number of linearly independent cointegrating vectors of the system. These vectors turn out to be the eigenvectors of a generalized eigenvalue problem when deriving the maximum likelihood estimator of the reduced VECM parameters. Johansen also developed a cointegration rank test for testing the one-sided hypothesis that the cointegration rank is smaller or equal than a given positive integer. The test depends on the so-called trace test statistic whose distribution has to be approximated by simulation. In this thesis, we have extended the table of the trace test statistic which was so far limited to relatively small systems.

The method introduced in this thesis, ICE, differs from that of Johansen in that instead of directly estimating a reduced VECM in which the cointegrating matrix appears explicitly, we first estimate a VAR which we rewrite as a VECM. Then, a low-rank approximation as well as a rank factorization of the so-called impact matrix is performed to retrieve the estimated cointegrating matrix. It can be seen as a two-step estimator with the first step consisting of a VAR estimation and the second step transforming the VAR estimates into VECM estimates and applying operations on the two-step estimate of the impact matrix. As such, it is more flexible than DiCE since the first step can be performed in numerous ways by choosing any VAR estimator.

Considering the recent developments in the ease of access to large data sets, we focus from a practical point of view on fairly large systems in this the-

sis, as one might encounter in applications. We are aware that there are far larger systems to deal with in applications, however we do not consider them for computational reasons. Consequently, we conjecture that in a fairly large cointegrated process, it might often be that several variables play no role at all in the cointegrating process or that the structure of time-dependence between variables in the system is sparse. In such a case, the use of regularized (or penalized) estimators would be preferred, which we choose to consider for the first step estimator of ICE in this thesis. The use of such regularized estimators is also motivated by the infeasibility in high-dimensional settings, i.e. the number of observations being smaller than the number of variables, of classical VAR estimators and Johansen's method for cointegration.

The remainder of this thesis is organized as follows. Since both Johansen's DiCE and ICE are heavily dependent on VAR processes, Section 2 gives a detailed background on their properties, various estimation methods and their limitations in high-dimensional cases. Section 3 introduces the cointegration framework of Johansen with the reduced form VECM. It also gives a detailed derivation of its (direct) cointegration estimator, as well as a proof of its infeasibility in high-dimensions. Johansen's trace statistic and cointegration rank test is also discussed. Section 4 presents a regularized VECM estimator introduced in [15], that we call the sparse DiCE (sDiCE). As mentioned in the above paragraph, three existing sparse (regularized) VAR estimators will be used for ICE, which we present in Section 5. We present our new method for cointegration estimation, ICE, in Section 6. Section 7 gives detail on the simulation study performed to assess the different methods' performances in different set-ups. Section 8 summarizes the results of that simulation study. We conclude this thesis in Section 9.

2 Vector Autoregressive Processes

We consider the k-dimensional Gaussian vector autoregressive process of the p-th order (VAR(p)) $\{X_t\}_{t=1,\dots,T}$ characterized by

$$X_t = \sum_{i=1}^{p} A_i X_{t-i} + \epsilon_t, \tag{1}$$

where $X_t = (X_{1,t}, \dots, X_{k,t})^\top$, $X_{-(p-1)}, \dots, X_0$ are deterministic, $A_i \in \mathbb{R}^{k\times k}$, $i = 1, 2, \dots, p$ are called the transition matrices containing the coefficient parameters and ϵ_t are *i.i.d.* $\mathcal{N}_k(0, \Omega)$ error terms with $\Omega \in \mathbb{R}^{k\times k}$ being symmetric and positive definite. Note that the process can also be expressed as

$$A(L)X_t = \epsilon_t \qquad \text{or} \tag{2}$$

$$X_t = \Phi^\top \Xi_{t-1} + \epsilon_t \tag{3}$$

with $A(L) = I_k - \sum_{i=1}^{p} A_i L^i$ being a matrix polynomial in L, which is defined as the lag operator $L^i X_t = X_{t-i}$, $\Phi^\top \coloneqq (A_1, A_2, \dots, A_p) \in \mathbb{R}^{k\times kp}$ and see (5) for Ξ_t.

To discuss the stability of this process, it is useful to consider its first order specification defined for $t \in [1, \infty)$ and proceed by recursion for $t = 1, 2, \dots$

$$\begin{aligned}
X_1 &= A_1 X_0 + \epsilon_1 \\
X_2 &= A_1(A_1 X_0 + \epsilon_1) + \epsilon_2 = A_1^2 X_0 + A_1\epsilon_1 + \epsilon_2 \\
&\vdots \\
X_j &= A_1^j X_0 + \sum\nolimits_{i=1}^{j-1} A_1^i \epsilon_i + \epsilon_j \\
&\vdots
\end{aligned}$$

This shows that X_j is only determined by the initial value X_0 and the sequence of errors $\{\epsilon_i\}_{i=1}^{j}$. Also, as $j \to \infty$, the process involves the infinite stochastic vector series $\sum_{i=1}^{\infty} A_1^i \epsilon_i$ which needs to converge in some sense for the process to be stable. From [6] and [10], we know that the series converges in mean square if $\{A_1^i\}$ is absolutely summable and if $\mathbb{E}\left[\epsilon_i^\top \epsilon_i\right] < \infty$. It is easy to see from the Jordan canonical form that a sufficient condition for the absolute summability is that all the moduli of the complex eigenvalues of A_1 are less that 1, since then $A_1^i \to 0$ as $i \to \infty$. This condition is equivalent to $\det(I_k - A_1 z) \neq 0 \ \forall\ |z| \leq 1,\ z \in \mathbb{C}$. The second condition is automatically fulfilled by assumption on the error term's covariance structure.

We can naturally generalize these results to the $p > 1$ case by rewriting the VAR(p) in companion form expressing another VAR(1) process

$$\Xi_t = A\Xi_{t-1} + \eta_t \tag{4}$$

where

$$\Xi_t = \underset{(kp\times 1)}{\begin{pmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-(p-1)} \end{pmatrix}}, \quad A = \underset{(kp\times kp)}{\begin{pmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I_k & 0 & \cdots & 0 & 0 \\ 0 & I_k & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_k & 0 \end{pmatrix}}, \quad \eta_t = \underset{(kp\times 1)}{\begin{pmatrix} \epsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}}. \tag{5}$$

We have that (4) is in fact

$$\begin{pmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-(p-1)} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{p} A_i X_{t-i} \\ X_{t-1} \\ \vdots \\ X_{t-(p-1)} \end{pmatrix} + \begin{pmatrix} \epsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Therefore, Ξ_t, and thus X_t, is stable if all the eigenvalues of A have a modulus less that 1, which as mentioned above is equivalent to $\det(I_{kp} - Az) = \det(A(z)) \neq 0 \ \forall \ |z| \leq 1$. We can also add that all the eigenvalues λ of A satisfy $\det\left(I_k\lambda^p - \sum_{i=1}^{p} A_i\lambda^{p-i}\right) = 0$ (see Appendix 10.A of [7]). All of these conditions are equivalently known as the stability condition.

At this stage, it is worthwhile to highlight that a VAR process that does not satisfy the stability condition is not necessarily undefined, though it might be explosive or integrated. For the remainder of this article, we will not assume by default that the stability condition is met.

2.1 Estimation of VAR Models

As we will see, we can express a VAR model as a general linear model (not to be mistaken with the generalized linear model or GLM) or simply a multivariate regression. Then we can either estimate the transition matrices by maximum likelihood thanks to the distributional assumption on the error terms or by generalized least squares (GLS). Alternatively, we can infer the distribution of each observation of the sample process conditioned on the past values and get an expression for the conditional joint density of the whole sample. Although these three estimators are identical, it is useful to review them since regularization techniques often involve variants of these three versions, which in general differ from each other unlike their unregularized peers.

2.1.1 Maximum Likelihood

Let $\{x_t\}_{t\in\mathcal{I}}$ where $\mathcal{I} = \{-(p-1), -(p-2), \ldots, T\}$ be an observed sample generated from (1). Denoting $c_{[a,b]} \coloneqq \{c_i\}_{i=a}^{b}$ and $\theta = \{\{A_i\}_{1\leq i\leq p}, \Omega\}$, we are interested in deriving the following conditional joint density

$$f_{X_{[1,T]}|X_{[-(p-1),0]}} \coloneqq f_{X_{[1,T]}|X_{[-(p-1),0]}}\left(x_{[1,T]} \mid x_{[-(p-1),0]}; \theta\right), \tag{6}$$

where $x_{[-(p-1),0]}$ are treated as the initial values on which the sample is conditioned, representing the conditional likelihood of the sample, which for conciseness we will call "likelihood" henceforth.

Since $\{x_t\}$ is deterministic (realized path of the generating process), by conditioning on all the values $\{x_i\}$ up to time t noninclusive, X_t is distributed as a constant plus $\epsilon_t \sim \mathcal{N}_k(0, \Omega)$, the only stochastic component, i.e.

$$X_t \mid x_{[-(p-1),t-1]} \sim \mathcal{N}_k(\Phi^\top \xi_{t-1}, \Omega), \tag{7}$$

where $\xi_t = (x_t, x_{t-1}, \ldots, x_{t-(p-1)})$. Therefore,

$$\begin{aligned} &f_{X_t|X_{[-(p-1),t-1]}} \\ &\quad = (2\pi)^{-k/2} \det(\Omega^{-1})^{1/2} \exp\left[-\frac{1}{2}\left(x_t - \Phi^\top \xi_{t-1}\right)^\top \Omega^{-1} \left(x_t - \Phi^\top \xi_{t-1}\right)\right]. \end{aligned} \tag{8}$$

We know that

$$f_{X_{[1,t]}|X_{[-(p-1),0]}} = f_{X_{[1,t-1]}|X_{[-(p-1),0]}} \cdot f_{X_t|X_{[-(p-1),t-1]}},$$

so by applying this recursively we can rewrite (6) as

$$f_{X_{[1,T]}|X_{[-(p-1),0]}} = \prod_{t=1}^{T} f_{X_t|X_{[-(p-1),t-1]}}, \tag{9}$$

in which substituting (8) and taking the logarithm yields the sample log likelihood

$$\begin{aligned} \mathcal{L}\left(x_{[1,t]};\, \theta\right) = &-\frac{Tk}{2}\log 2\pi + \frac{T}{2}\log\left(\det\left(\Omega^{-1}\right)\right) \\ &-\frac{1}{2}\sum_{t=1}^{T}\left(x_t - \Phi^\top \xi_{t-1}\right)^\top \Omega^{-1}\left(x_t - \Phi^\top \xi_{t-1}\right). \end{aligned} \tag{10}$$

We find the maximum likelihood estimator (MLE) of $\Phi^\top$ by solving the equation that sets the differentiated log likelihood (10) w.r.t. $\Phi^\top$ to zero, which yields

$$\widehat{\Phi}^\top = \left(\sum_{t=1}^{T} x_t \xi_{t-1}^\top\right)\left(\sum_{t=1}^{T} \xi_{t-1}\xi_{t-1}^\top\right)^{-1}. \tag{11}$$

To find the MLE of Ω, we first substitute (11) in (10)

$$\mathcal{L}\left(x_{[1,t]};\, \widehat{\Phi}, \Omega\right) = -\frac{Tk}{2}\log 2\pi + \frac{T}{2}\log\left(\det\left(\Omega^{-1}\right)\right) - \frac{1}{2}\sum_{t=1}^{T} \widehat{\epsilon_t}^\top \Omega^{-1} \widehat{\epsilon_t},$$

where $\widehat{\epsilon_t} \coloneqq x_t - \widehat{\Phi}^\top \xi_{t-1}$ and differentiated w.r.t. Ω^{-1} to get

$$\widehat{\Omega} = T^{-1}\sum_{t=1}^{T} \widehat{\epsilon_t}\widehat{\epsilon_t}^\top. \tag{12}$$

2.1.2 General Linear Model

We can easily view (1) as a general linear model (GenLM) in which each of the k dependent variables is regressed on the same kp independent variables. The dependent variables are $\{X_{i,t}\}_{1\leq i\leq k}$ and the independent variables are $\{X_{i,t-j}\}_{1\leq i\leq k, 1\leq j\leq p}$. If we write a sample drawn from (3) in matrix notation, i.e. we stack the transpose of x_t for $t = 1, \ldots, T$, we obtain

$$Y = X\Phi + E \tag{13}$$

where

$$Y = \underset{T\times k}{\begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_T^\top \end{pmatrix}}, \quad X = \underset{T\times kp}{\begin{pmatrix} \xi_0^\top \\ \xi_1^\top \\ \vdots \\ \xi_{T-1}^\top \end{pmatrix}} \quad \text{and} \quad E = \underset{T\times k}{\begin{pmatrix} e_1^\top \\ e_2^\top \\ \vdots \\ e_T^\top \end{pmatrix}}, \tag{14}$$

and ξ_t is an observation drawn from Ξ_t.

Note that the coefficient matrix Φ in the matrix notation (13) is simply the transpose of that in the observational notation, $\Phi^\top$ (3).

MLE Since each row of E is drawn from $\epsilon_t^\top \sim$ i.i.d. $\mathcal{N}_k(0, \Omega)$, it follows that $x_t^\top - \xi_{t-1}\Phi \sim$ i.i.d. $\mathcal{N}_k(0, \Omega)$. Thus, we know that the likelihood of the sample Y is the likelihood of an i.i.d. multivariate Gaussian sample (as in Section 2.1.1)

$$\begin{aligned} \mathcal{L}(Y;\, \Phi, \Omega) = &-\frac{Tk}{2}\log 2\pi + \frac{T}{2}\log\left(\det\left(\Omega^{-1}\right)\right) \\ &-\frac{1}{2}\operatorname{tr}\left[(Y - X\Phi)\Omega^{-1}(Y - X\Phi)^\top\right], \end{aligned} \tag{15}$$

since $\sum_{t=1}^T \epsilon_t^\top \Omega^{-1}\epsilon_t = \operatorname{tr}\left(\Omega^{-1}\sum_{t=1}^T \epsilon_t\epsilon_t^\top\right) = \operatorname{tr}\left(\Omega^{-1}E^\top E\right) = \operatorname{tr}\left(E^\top\Omega^{-1}E\right)$. The first and third equality signs come from the property of the trace (see A.2.2 from [11]) and the second follows from the fact that the rows of E are $\epsilon_t^\top$ (in the same fashion as in Appendix A). Now, the MLEs from the GenLM of Φ and Ω are given by

$$\Phi^* = \left(X^\top X\right)^{-1} X^\top Y \tag{16}$$

$$\Omega^* = T^{-1}Y^\top PY \tag{17}$$

with the projection matrix $P = I_T - X\left(X^\top X\right)^{-1}X^\top$, which effectively resemble to the OLS estimators of a multiple linear regression, in which case Y would be a vector instead of a matrix. We show in Appendix A that $(\Phi^*)^\top = \widehat{\Phi}^\top$ and $\widehat{\Omega} = \Omega^*$ in Appendix B.

GLS Consider the model (13) written in vectorized form as

$$Y_V = \widetilde{X}\Phi_V + E_V \tag{18}$$

where $M_V \coloneqq \text{vec}(M) \in \mathbb{R}^{mn}$, $M \in \mathbb{R}^{m \times n}$, $\widetilde{X} \coloneqq I_k \otimes X \in \mathbb{R}^{Tk \times k^2 p}$ and $E^V \sim \mathcal{N}_{k^2}(0, \Sigma)$, $\Sigma = \Omega \otimes I_T$. In that case,

$$\begin{aligned}
\widehat{\Phi}_V &= \underset{\Phi_V}{\arg\min} \left(Y_V - \widetilde{X}\Phi_V\right)^\top \Sigma^{-1} \left(Y_V - \widetilde{X}\Phi_V\right) \\
&= \left(\widetilde{X}^\top \Sigma^{-1} \widetilde{X}\right)^{-1} \widetilde{X}^\top \Sigma^{-1} Y_V \\
&= \left[I_k \otimes \left(X^\top X\right)^{-1} X^\top\right] Y_V. \qquad (19)
\end{aligned}$$

Again, we see that $\text{vec}(\Phi^*) = \text{vec}(\widehat{\Phi}) = \widehat{\Phi}_V$, showing the equivalence between these three estimation frameworks, stemming from the normality assumption on the error terms.

2.2 The High-Dimensional Setting

All of the above estimation methods require the inversion of the $kp \times kp$ moment matrix $X^\top X$. However, this matrix might not always be invertible and it could be singular for a variety of reasons. Since it is a product of the same $T \times kp$ matrix X, the design matrix, we can infer bounds for the rank of the moment matrix from the rank bounds of the design matrix. Since our focus here is on the upper bound (simplest case), we need only look the dimension of X and discriminate between two cases. Here we show that $T < kp$ is a sufficient condition for $X^\top X$ to be singular. On the contrary, if $T \geq kp$, $X^\top X$ is singular only in the special case where the columns of the design matrix are not all linearly independent.

Proof.

Case 1: $T < kp \Rightarrow \text{rank}\left(X^\top X\right) = \text{rank}(X) \leq T < kp \Rightarrow \det\left(X^\top X\right) = 0$

Case 2: $T \geq kp \Rightarrow \text{rank}\left(X^\top X\right) = \text{rank}(X) \leq kp \Rightarrow \det\left(X^\top X\right) = 0$
$\iff \dim\left(\text{colsp}(X)\right) = \text{rank}(X) < kp$

■

This shows that the classical estimation methods presented above are infeasible when $kp > T$, i.e. in the high-dimensional setting. It is this realization, in part, that motivates the use of regularization in estimation problems. Indeed, regularization serves two purposes. The first purpose is computational and in this case relates with the above-mentioned problem, where an unregularized estimator is unfeasible but a regularized one is often feasible. The second purpose of regularization is statistical in the sense that it allows to construct estimators that might better exploit subjacent low-dimensional or sparse structure in the true underlying data generating process, which lead to desirable properties of estimators. The focus of this thesis does not lie on the derivation of these properties but rather highlighting the clear computational advantage of regularized estimators in terms of feasibility as well as their capability of uncovering

sparse underlying parametric structures, in turn preventing overfitting and thus yielding better forecasting performances.

In order to better grasp the idea of regularization, we briefly introduce the broad class of regularized M-estimators [14]. We want to estimate a parameter $\theta_0 \in \mathbb{R}^q$ by minimizing the average of a loss function and a regularizer. Formally, let $\mathcal{Z}^n \coloneqq \{z_1, \dots, z_n\}$ be a set of observations drawn from Z with marginal distribution $\mathbb{P}$ and $\theta_0 = \theta_0(\mathbb{P})$. An M-estimator is defined as

$$\widehat{\theta} \in \underset{\theta \in \mathbb{R}^q}{\arg\min}\, \ell_n(\theta, \mathcal{Z}^n) \coloneqq n^{-1} \sum_{i=1}^{n} \ell_n(\theta, z_i), \tag{20}$$

where $\ell_n : \mathbb{R}^q \times \mathcal{Z} \to \mathbb{R}$ is a loss function usually chosen such that

$$\theta_0 = \underset{\theta \in \mathbb{R}^q}{\arg\min}\, \mathbb{E}[\ell_n(\theta, Z)].$$

A regularized M-estimator combines (20) with a regularizer or penalty $\rho : \mathbb{R}^q \to \mathbb{R}^+$ which is what imposes a certain structure to the estimates. The regularizer can be included in two ways; either as a constraint imposed to (20) or with a Lagrange multiplier in (20):

$$\widehat{\theta} \in \underset{\theta \in \mathbb{R}^q}{\arg\min}\, \ell_n(\theta, \mathcal{Z}^n) \quad \text{s.t.} \quad \rho(\theta) \leq r \qquad \text{or} \tag{21}$$

$$\widehat{\theta} \in \underset{\theta \in \mathbb{R}^q}{\arg\min}\, \ell_n(\theta, \mathcal{Z}^n) + \lambda \rho(\theta), \tag{22}$$

with $r, \lambda > 0$. Often however, no analytical solution to (21) or (22) exists rather these must be solved numerically. A case that can be solved analytically, though, is the ridge estimator. Its multivariate extension [2] applied to the model (18) is defined as

$$\widehat{\Phi}_R(\lambda) \in \underset{\Phi^V \in \mathbb{R}^{k^2 p}}{\arg\min} \left\| Y^V - \widetilde{X} \Phi^V \right\|_2^2 + \lambda \left\| \Phi^V \right\|_2^2 \tag{23}$$

where the loss function corresponds to the much used squared L^2-norm of the residuals E^V and the regularizer is a penalty on the squared L^2-norm of the parameter Φ^V. This estimator has the following solution

$$\widehat{\Phi}_R(\lambda) = \left(I_k \otimes X^\top X + \lambda I_{k^2 p} \right)^{-1} \left(I_k \otimes X^\top \right) Y^V. \tag{24}$$

Thus, contrary to the estimators in Section 2.1, (24) is usually feasible even when $kp > T$. Indeed, the matrix to be inverted is no longer $X^\top X$ but $M_R \coloneqq I_j \otimes X^\top X + \lambda I_{k^2 p}$. Assuming that $\operatorname{rank}(X^\top X) = r < kp$, $\operatorname{rank}(I_j \otimes X^\top X) = kr$ thus $\operatorname{rank}(M_R) \leq k^2 p$ and $\det(M_R) \neq 0 \iff \operatorname{rank}(M_R) = k^2 p$.

3 Johansen's Cointegration Framework

The theory developed by Søren Johansen to estimate cointegrating relationships is based on the estimation of a reduced form vector error correction model (VECM), which is an equivalent way of writing a VAR process. The VECM form equivalent of a VAR(p) process (1) is defined as

$$\Delta X_t = \Pi X_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta X_{t-i} + \epsilon_t, \tag{25}$$

where $\Pi = -I_k + \sum_{i=1}^{p} A_i$ and $\Gamma_i = -\sum_{j=i+1}^{p} A_j$.

Cointegration is then formulated as a condition on the (reduced) rank of Π, the impact matrix. If $\text{rank}(\Pi) = r < k$, it can be rewritten as

$$\Pi = \alpha\beta^\top, \tag{26}$$

where $\alpha, \beta \in \mathbb{R}^{k\times r}$. The unrestricted VECM (25) becomes the reduced form VECM

$$\Delta X_t = \alpha\beta^\top X_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta X_{t-i} + \epsilon_t. \tag{27}$$

The cointegrating vectors are the columns of β, meaning that $\beta^\top X_t$ is an r-dimensional covariance stationary process conditional on a suitable choice of the initial values of X_t. To see why β is a cointegrating matrix, we briefly state Theorem 4.2 of [9], also known as the Johansen-Granger representation theorem, which first needs some notation. For any $m \in \mathbb{R}^{l\times n}$, let $m_\perp \in \mathbb{R}^{l\times(l-n)}$ with $\text{rank}(m_\perp) = l - n$ be such that $m^\top m_\perp = 0$. Also, we call the characteristic polynomial of (25) $A'(z) = (1-z)I_k - \Pi z - \sum_{i=1}^{p-1}\Gamma_i(1-z)z^i$ and let $\Gamma = I_k - \sum_{i=1}^{p}\Gamma_i$, $C = \beta_\perp(\alpha_\perp^\top\Gamma\beta_\perp)^{-1}\alpha_\perp^\top$ and $P_m = m(m^\top m)^{-1}m^\top$. The theorem states that if $\det(A(z)) = 0 \Rightarrow |z| > 1$ or $z = 1$, and $\text{rank}(\Pi) = r < k$, then X_t can be represented as

$$X_t = C\sum_{i=1}^{t}\epsilon_i + C(L)\epsilon_t + P_{\beta_\perp}X_0 \tag{28}$$

if and only if $\text{rank}(\alpha_\perp^\top\Gamma\beta_\perp) = k - r$. In this case, (28) describes a cointegrated $I(1)$ process since both ΔX_t and $\beta^\top X_t$ can be made stationary given a suitable choice of the initial value X_0. Indeed, X_t contains both a stochastic trend and a stationary component. The stochastic trend $C\sum_{i=1}^{t}\epsilon_i$ becomes stationary after differencing once ($C\epsilon_t$) or completely disappears by multiplying by $\beta^\top$ ($\beta^\top C = 0$). The stationary component is $C(L)\epsilon_t$, where $C(z) = \sum_{i=0}^{\infty} C_i z^i$, with $C_i \in \mathbb{R}^{k\times k}$, converges for $|z| \leq 1 + \delta$ for some $\delta > 0$.

The $I(1)$ model (25) is called H(r) for $r = 0, \ldots, k$ such as it expresses the hypothesis that $\text{rank}(\Pi) \leq r$. Consequently, the H(r) models are a nested sequence of models

$$\text{H}(0) \subset \cdots \subset \text{H}(r) \subset \cdots \subset \text{H}(k). \tag{29}$$

H(0) corresponds to the restriction $\Pi = 0$, i.e. the non-cointegrated $I(1)$ since ΔX_t is an $I(0)$ VAR$(p-1)$ and H(k) represents the unrestricted VAR since it is just a VAR rewritten in VECM. Thus, H(r) for $r = 1, \dots, k-1$ are the models that guarantee cointegration.

It is important to note that the cointegrating vectors β and the adjustment coefficients α (forces pushing the cointegrated relationships back to equilibrium) are not uniquely identified, since for any non-singular $m \in \mathbb{R}^{r \times r}$ we can find $\alpha' = \alpha m^{-1}$ and $\beta' = m\beta$ such that $\Pi' = \alpha'\beta' = \Pi$.

3.1 Johansen's Direct Cointegration Estimation

The goal of this section is to derive the direct cointegration estimator (DiCE) developed by Johansen to estimate in a direct manner the cointegrating vectors β for a given cointegrating rank r, as well as its one-tailed test statistic for the hypothesis on the latter. To that end, we introduce the following notation. Let $Z_{0,t} = \Delta X_t$, $Z_{1,t} = X_{t-1}$, $Z_{2,t} = \text{vec}(\Delta X_{t-1}, \dots, \Delta X_{t-(p-1)})$ and $G = (\Gamma_1, \dots, \Gamma_{p-1})$ such that (27) can be written as

$$Z_{0,t} = \alpha\beta^\top Z_{1,t} + GZ_{2,t} + \epsilon_t. \tag{30}$$

Also, for any $m_{i,t} \in \mathbb{R}^{d_i}$, we denote the corresponding product-moment matrices by $M_{ij} = T^{-1}\sum_{t=1}^{T} m_{i,t}m_{j,t}^\top$, noting that $M_{ij} = M_{ji}^\top$.

As in (15), the log-likelihood of the VECM sample $\{z_{0,t}\}$, aside from the first constant, is straightforwardly given by

$$\begin{aligned}\log \mathcal{L}(\alpha, \beta, G, \Omega) = \frac{T}{2}\log\left[\det\left(\Omega^{-1}\right)\right] \\ - \frac{1}{2}\sum_{t=1}^{T}(z_{0,t} - \alpha\beta^\top z_{1,t} - Gz_{2,t})^\top\Omega^{-1}(z_{0,t} - \alpha\beta^\top z_{1,t} - Gz_{2,t}).\end{aligned} \tag{31}$$

The autoregressive coefficients G can at this stage only be estimated conditionally on the cointegration parameters α and β, with the first order conditions

$$\sum_{t=1}^{T}(z_{0,t} - \alpha\beta^\top z_{1,t} - \widehat{G}z_{2,t})z_{2,t}^\top = 0$$

yielding the estimator

$$\widehat{G}(\alpha, \beta) = M_{02}M_{22}^{-1} - \alpha\beta^\top M_{12}M_{22}^{-1}. \tag{32}$$

Since the transition matrices are not the main parameter of interest in this problem, we wish to profile out G to obtain the profile likelihood, function of the other parameters. To do so, we use the residuals that we would have obtained by regressing both $Z_{0,t}$ and $Z_{1,t}$ on $Z_{2,t}$. Written as a GenLM, we get

$$\begin{aligned} Z_0 &= Z_2\theta_0^\top + E_0 \\ \text{and} \quad Z_1 &= Z_2\theta_1^\top + E_1, \end{aligned}$$

where $Z_i = (Z_{i,1}, \ldots, Z_{i,T})^\top$, $i = 0, 1, 2$. Thus, from (16), we know that for $i = 0, 1$

$$\widehat{\theta}_i^\top = \left(Z_2^\top Z_2\right)^{-1} Z_2^\top Z_i = M_{22}^{-1} M_{2i},$$

such that in the observational notation we get

$$\widehat{\theta}_i = M_{i2} M_{22}^{-1}.$$

This leads to the following residuals in the observational notation

$$\begin{aligned} R_{0,t} &= Z_{0,t} - \widehat{\theta}_0 Z_{2,t} = Z_{0,t} - M_{02} M_{22}^{-1} Z_{2,t} \\ R_{1,t} &= Z_{1,t} - \widehat{\theta}_1 Z_{2,t} = Z_{1,t} - M_{12} M_{22}^{-1} Z_{2,t} \end{aligned}$$

with which we can write the profile likelihood equivalent to (31)

$$\begin{aligned} \log \mathcal{L}(\alpha, \beta, \Omega) = \frac{T}{2} \log \left[\det\left(\Omega^{-1}\right)\right] \\ - \frac{1}{2} \sum_{t=1}^{T} \left(R_{0,t} - \alpha\beta^\top R_{1,t}\right)^\top \Omega^{-1} \left(R_{0,t} - \alpha\beta^\top R_{1,t}\right). \end{aligned} \tag{33}$$

This would be the likelihood of $\{\tilde{\epsilon}_t\}$ in

$$R_{0,t} = \alpha\beta^\top R_{1,t} + \tilde{\epsilon}_t$$

which is effectively a reduced rank regression. Then we can estimate α and Ω conditional on β by regressing $R_{0,t}$ on $\beta^\top R_{1,t}$, thereby obtaining

$$\widehat{\alpha}(\beta) = S_{01}\beta \left(\beta^\top S_{11}\beta\right)^{-1} \tag{34}$$

$$\begin{aligned} \widehat{\Omega}(\beta) &= S_{00} - S_{01}\beta \left(\beta^\top S_{11}\beta\right)^{-1} \beta^\top S_{10} \\ &= S_{00} - \widehat{\alpha}(\beta) \left(\beta^\top S_{11}\beta\right) \widehat{\alpha}(\beta)^\top, \end{aligned} \tag{35}$$

where $S_{ij} = T^{-1} \sum_{t=1}^{T} R_{i,t} R_{j,t}^\top = M_{ij} - M_{i2} M_{22}^{-1} M_{2j}$ for $i, j = 0, 1$. Now, substituting (34) and (35) into (33), we obtain the maximized likelihood (leaving out the first constant and taking the exponential) solely depending on β

$$\mathcal{L}_{max}^{-2/T}(\beta) = \det\left(\widehat{\Omega}(\beta)\right) = \det\left(S_{00} - S_{01}\beta \left(\beta^\top S_{11}\beta\right)^{-1} \beta^\top S_{10}\right). \tag{36}$$

To maximize (36) with respect to β, we first rewrite it using an identity on the determinant of block matrices (see Theorem A.3.2 of [1])

$$\begin{aligned} \det\left(S_{00} - S_{01}\beta \left(\beta^\top S_{11}\beta\right)^{-1} \beta^\top S_{10}\right) \\ = \det(S_{00}) \frac{\det\left(\beta^\top \left(S_{11} - S_{10} S_{00}^{-1} S_{01}\right)\beta\right)}{\det\left(\beta^\top S_{11}\beta\right)}. \end{aligned} \tag{37}$$

Therefore, maximizing (36) amounts to maximizing the right-hand side fraction of (37), which is of the form

$$h(x) = \det\left(x^\top M x\right) / \det\left(x^\top N x\right) \tag{38}$$

with $x \in \mathbb{R}^{k\times r}$ and $M, N \in \mathbb{R}^{k\times k}$ both symmetric, positive semi-definite and positive definite, respectively. By Lemma A.8 of [9], the maximum of (38) is reached for $\widehat{x} = (v_1, \ldots, v_r)$ with $h(\widehat{x}) = \prod_{i=1}^{r} \lambda_i$ where λ_i and v_i are solutions to the following generalized eigenvalue problem

$$\det(\lambda N - M) = 0, \tag{39}$$

furthermore assuming that the eigenvectors v_i in $\widehat{x}$ are arranged such that their associated eigenvalues λ_i are in decreasing order; $1 > \lambda_1 > \ldots > \lambda_r > 0$.

In our case, we need to solve the following generalized eigenvalue problem

$$\begin{aligned} &\det\left(\rho S_{11} - \left(S_{11} - S_{10}S_{00}^{-1}S_{01}\right)\right) = 0 \\ \iff &\det\left(\lambda S_{11} - S_{10}S_{00}^{-1}S_{01}\right) = 0, \end{aligned} \tag{40}$$

where $\lambda = 1 - \rho$ and $\hat{v}_i$ are the eigenvectors associated to $\hat{\lambda}_i$ for $i = 1, \ldots, k$. Therefore, under the hypothesis H(r), the Johansen's estimator of β is

$$\widehat{\beta} = (\hat{v}_1, \ldots, \hat{v}_r), \tag{41}$$

for which the maximized likelihood takes value

$$\mathcal{L}_{max}^{-2/T}(r) = \det(S_{00}) \prod_{i=0}^{r} (1 - \hat{\lambda}_i). \tag{42}$$

Note that all the models H(r) for $r = 0, \ldots, k$ are solved by the same generalized eigenvalue problem. Indeed, all the eigenvectors and eigenvalues are computed once from one problem and then eigenvectors are added as cointegrating vectors in decreasing order of their corresponding eigenvalue. In fact, this is motivated by the fact that these eigenvalues are equal to the squared canonical correlations between $R_{0,t}$ and $R_{1,t}$. This implies that the r first columns of β are such that the linear combination process $\beta^\top X_{t-1}$ is the most correlated with the differentiated process ΔX_t, after correcting for the lags $\Delta X_{t-1}, \ldots, \Delta X_{t-(p-1)}$. Therefore, Johansen's DiCE method aims at finding the cointegrating relationships that are the most likely to be stationary, with this likelihood decreasing in the columns of the cointegrating matrix.

We can naturally obtain a likelihood ratio test statistic for H(r) by dividing (42) by the same function but evaluated at $r = k$

$$Q\left(\mathrm{H}(r) \mid \mathrm{H}(k)\right)^{-2/T} = \frac{\det(S_{00}) \prod_{i=1}^{r} (1 - \hat{\lambda}_i)}{\det(S_{00}) \prod_{i=1}^{k} (1 - \hat{\lambda}_i)},$$

which leads to Johansen's trace statistic

$$\tau(r \mid k) = -2 \log Q\left(H(r) \mid H(k)\right) = -T \sum_{i=r+1}^{k} \log(1 - \hat{\lambda}_i). \tag{43}$$

The trace statistic weakly converges to the trace of the product of matrix stochastic integrals

$$\tau(r \mid k) \xrightarrow{d} \operatorname{tr}\left[\left(\int_0^1 dB\right) B^\top \left(\int_0^1 B_u B_u^\top du\right) \int_0^1 B dB^\top\right], \tag{44}$$

where B is a $(k-r)$-dimensional Brownian motion.

3.2 Infeasibility in High Dimensions

As we know from Section 2.2, the infeasibility of an estimator can arise from the singularity of matrices needed in the computation of the estimator. In the high-dimensional setting, i.e. when the number of variables exceeds the number of available observations, in our case $kp > T$, the structure of the data has a direct impact on the rank of certain matrices that need to be inverted in the derivation of an estimator.

In Johansen's DiCE problem, there are three square matrices that need to be inverted to compute the estimators of G, α, Ω as well as of β; namely, M_{22}, $\beta^\top S_{11}\beta$ and S_{00}. However, the latter two matrices both depend on M_{22}. Thus, if M_{22} is singular, none of the estimators can be computed. Here, we show that Johansen's estimation method is even more restrictive in terms of the feasibility space of the dimensionality of the process. Indeed, the k-dimensional VAR(p) estimation is not feasible if $T < kp$, whereas Johansen's estimation is not feasible if $T < k(p-1)$.

Recall that $M_{22} = T^{-1}\sum_{t=1}^{T} Z_{2,t} Z_{2,t}^\top$ where $Z_{2,t} \in \mathbb{R}^{k(p-1)}$. Thus, M_{22} involves a sum of outer products of the same vectors. This implies that $\operatorname{rank}\left(Z_{2,t} Z_{2,t}^\top\right) = 1$, which means that $\operatorname{rank}\left(\sum_{t=1}^{T} Z_{2,t} Z_{2,t}^\top\right) = \operatorname{rank}(M_{22}) \leq T$. However, $M_{22} \in \mathbb{R}^{k(p-1)\times k(p-1)}$, such that $T < k(p-1)$ automatically implies that $\det(M_{22}) = 0$. ■

3.3 Trace Statistic Limit Distribution Approximation

Since the limit random variable in (44)

$$W(k-r) = \operatorname{tr}\left[\left(\int_0^1 dB\right) B^\top \left(\int_0^1 B_u B_u^\top du\right) \int_0^1 B dB^\top\right] \tag{45}$$

is a complicated function of Brownian motions, an analytical distribution does not seem to possibly be derived. Instead, we need to find another discrete approximation that also weakly converges to $W(k-r)$. Then, we use simulation to draw an empirical distribution of that new random variable, which in turn will serve as an approximation for the true distribution of $W(k-r)$.

As Johansen pointed out, the distribution of $W(k-r)$ mainly depends on the dimension of the problem, namely the number of possibly non-stationary components k of the system. As it is implicitly summed up in (44), the limit

distribution of $\tau(r \mid k)$ is the same as that of $\tau(0 \mid k-r)$. That is, testing $\mathrm{H}(r) \subset \mathrm{H}(k)$ amounts to testing $\mathrm{H}(0) \subset \mathrm{H}(k-r)$, meaning testing $\Pi = 0$ in

$$\Delta X_t = \Pi X_{t-1} + \epsilon_t$$

where $\Pi \in \mathbb{R}^{(k-r)\times(k-r)}$. Thus, under the null, the $(k-r)$-dimensional system is described by $\Delta X_t = \epsilon_t$.

We approximate $W(k-r)$ by

$$\widehat{W}(k-r) = \operatorname{tr}\left[\sum_{t=1}^{T} \epsilon_t X_{t-1}^\top \left(\sum_{t=1}^{T} X_{t-1} X_{t-1}^\top\right)^{-1} \sum_{t-1}^{T} X_{t-1}\epsilon_t^\top\right], \tag{46}$$

where $\epsilon_t \sim$ i.i.d. $\mathcal{N}_{k-r}(0, I_{k-r})$ and $X_t = X_0 + \sum_{i=1}^{t} \epsilon_i$ with $X_0 = 0$. For our simulation, we choose $T = 2000$ for $k - r = 1, \ldots, 200$ and proceed to 5000 iterations. Prior simulations that we are aware of all used smaller sample sizes and iterations, with the highest dimension $k-r$ only reaching 11 [13]. Our tabulated quantiles resulting from this simulation can be found on `https://jonkq.github.io/research`.

3.4 Cointegrating Rank Test and Estimator

We can use this table to test hypotheses on $\mathrm{H}(r)$. From (43), we see that the trace statistic is a decreasing function of r for fixed $\hat{\lambda}_i$, meaning for a given realized process. This can also be observed from the quantiles table, where for a fixed k, rows further down the table correspond to smaller ranks r. More importantly, recall from (29) that a model $\mathrm{H}(r)$ contains all antecedent models $\mathrm{H}(0), \ldots, \mathrm{H}(r-1)$, since $\mathrm{H}(r)$ expresses the one-tailed hypothesis $\operatorname{rank}(\Pi) \leq r$. Thus, we can test the null hypothesis $\operatorname{rank}(\Pi) \leq r$ against the alternative hypothesis $\operatorname{rank}(\Pi) > r$. For a significance level $0 < \alpha < 1$, we reject the null if $\tau(r \mid k) > c_{1-\alpha}(k-r)$, where the latter represents the $100(1-\alpha)$-th percentile of $\widehat{W}(k-r)$. If we have an informative prior on r, we can proceed exactly this way to test it. Without any prior, we can use this test successively to construct $\hat{r}$, an estimator of r, as such

$\hat{r} \leftarrow 0$
while $\tau(\hat{r} \mid k) > c_{1-\alpha}(k-\hat{r})$ **do**
  $\hat{r} \leftarrow \hat{r} + 1$
end while

where in fact $\mathrm{H}(r)$ is tested for increasing r as we reject it until we cannot reject it anymore and $\hat{r}$ is left at its last value.

4 Sparse Direct Cointegration Estimation

The method developed in [15] aims at finding a sparse and direct cointegration estimator based on the VECM. The estimator is sparse in the sense that the L^0-norm of the cointegrating vectors (columns of β) is reduced compared to Johansen's DiCE. It also solves the problem of high-dimensionality rendering Johansen's technique impossible, since it uses regularization.

Consider the VECM in (30) written in matrix notation

$$Z_0 = Z_1\Pi^\top + Z_2 G^\top + E \tag{47}$$

where $Z_i = (Z_{i,1}, \ldots, Z_{i,T})^\top,\ i = 0, 1, 2$ and $E = (\epsilon_1, \ldots, \epsilon_T)^\top$. The negative log-likelihood of (47) is given by (see (15))

$$\begin{aligned} -\mathcal{L}(Z_0; \Pi, G, \Omega) = &-\frac{1}{2}\log(\det(\Omega^{-1})) \\ &+ \frac{1}{2T}\operatorname{tr}\left[(Z_0 - Z_1\Pi^\top - Z_2 G^\top)\Omega^{-1}(Z_0 - Z_1\Pi^\top - Z_2 G^\top)^\top\right]. \end{aligned} \tag{48}$$

Consider now adding a regularizer to each β, G and Ω to (48) such as

$$Q_{\mathrm{s}}(\Pi, G, \Omega, \lambda) = -\mathcal{L}(Z_0; \Pi, G, \Omega) + (\|\beta\|_{1,1}, \|G\|_{1,1}, \|\Omega\|_{1,1})\lambda, \tag{49}$$

with $\lambda \in \mathbb{R}^3$ and $\lambda_i > 0,\ i = 1, 2, 3$.

The optimization problem thus consists of minimizing (49) with respect to β, G and Ω, subject to the constraint $\Pi = \alpha\beta^\top$. Since three estimators must be found, including the sparse DiCE of β, we must decompose this optimization into three optimizations and iteratively solve and substitute each of them into the others (see [15] for the detailed algorithm).

5 Sparse VAR Estimation Methods

This section presents some methods that use regularization in the estimation of VAR models. Once again, the objective of these methods is twofold. As we have seen in Section 2.2, regularization is required in the high-dimensional cases just to produce feasible estimators, or so to say regularize ill-posed inference problems. Additionally, as we will see, regularization generally induces shrinkage, in the sense that the resulting coefficients get closer to zero or equivalently that the effect of variables on other variables is reduced. This second feature of regularization has the advantage that the variance of a shrinkage estimator is usually diminished, though bias might be introduced. The ridge estimator in (23) is a typical example of such estimators.

In this thesis though, our interest lies on regularized estimator that induce sparsity, which goes further than mere shrinkage. This is motivated by the hypothesis of underlying parsimonious interdependence structures in very large systems. In other words, each dependent variable is only affected by a possibly different subset of all independent variables. Of course, in inference problems, a very small estimated coefficient might probably not be significantly different from zero in reality. Thus, forcing such coefficients to be exactly zero might result in improved prediction power.

In the case of VAR models, there are k independent variables, the individual processes in the system, and kp dependent variables, the lagged processes. Thus, a VAR model contains k^2p coefficients, the number of elements in Φ. If we let a given sparse VAR (sVAR) model $\mathcal{M}_n$ be characterized solely by the set of pairs of row and column of $\Phi^\top$ corresponding to the non-zero coefficients of $\mathcal{M}_n$, then this set can be denoted by $\mathcal{S}_n \subseteq \mathcal{S} = \{1, \dots, k\} \times \{1, \dots, kp\}$. We know that $|\mathcal{P}(\mathcal{S})| = 2^{k^2p}$, which therefore represents the number of possible sVAR models $\mathcal{M}_n,\ n = 1, \dots, 2^{k^2p}$.

This shows that there is a computationally prohibitive number of sVAR models to estimate even for moderate k and p if we were to select the best one based on information criteria. Other methods including forward or backward selection or those involving hypothesis tests on coefficients might prove to be computationally less demanding, though they are rarely optimal in selecting the best model. Therefore, another advantage of sparse estimators using regularization is that they perform estimation as well as model selection simultaneously, resulting in computational gains. We present three different existing methods

5.1 Lasso-VAR

One way of applying lasso to a VAR model is to consider the vectorized form (18) [12]. The lasso estimator $\Phi_V^{\ell_1}$ is then obtained by minimizing the sum of squared residuals

$$\left\| Y_V - \widetilde{X}\Phi_V \right\|_2^2$$

with respect to Φ_V, subject to the constraint

$$\|\Phi_V\|_1 \leq s,\ s > 0.$$

This estimator resulting from this constrained minimizing problem can be equivalently solved as follows

$$\Phi_V^{\ell_1} = \underset{v \in \mathbb{R}^{k^2 p}}{\arg\min}\ Q_{\ell_1}(v, \lambda), \tag{50}$$

where

$$Q_{\ell_1}(v, \lambda) = \left\| Y_V - \widetilde{X} v \right\|_2^2 + \lambda \|v\|_1 \tag{51}$$

for a certain value of the tuning parameter λ depending on the other tuning parameter s. Note that for $\lambda = 0$ or for a sufficiently large s, the lasso estimator is identical to the classical least square estimate (19), or $\Phi_V^{\ell_1} = \widehat{\Phi}_V$. As λ increases though, coefficients will be shrunk towards zero, with the least significant ones being forced to exactly zero. As the full model is obtained with $\lambda = 0$, we obtain the empty model, in which all coefficients equal zero, for a sufficiently large λ. The lasso estimator is usually solved by means of the Least Angle Regression (LARS) algorithm (see [4]).

5.2 Regularized Yule-Walker

5.2.1 Yule-Walker Setup

Unlike other estimation methods presented so far, this method introduced in [8] is not based on neither least square nor maximum likelihood problems. Instead, it exploits the Yule-Walker equation for VAR(1) processes which implies that a certain matrix is restricted. Indeed, consider (1) with $p = 1$ for which the stability condition is satisfied such that the process is stationary. This implies that $X_t \sim \mathcal{N}_k(0, \Sigma_X)$ and $\mathbb{E}[X_t X_{t-i}^\top] = Cov(X_t, X_{t-i}) \coloneqq \Sigma_X^{(i)}$. We can calculate its i-th lag autocovariance by multiplying by X_{t-i} and taking expectations

$$\begin{aligned}
\mathbb{E}[X_t X_{t-i}^\top] &= A_1 \mathbb{E}[X_{t-1} X_{t-i}^\top] \\
&= A_1 \mathbb{E}[(A_1 X_{t-2} + \epsilon_{t-1}) X_{t-i}^\top] \\
&= A_1^2 \mathbb{E}[X_{t-2} X_{t-i}^\top] \\
&\vdots \\
&= A_1^i \mathbb{E}[X_{t-i} X_{t-i}^\top],
\end{aligned}$$

since $\mathbb{E}[X_{t-i} \epsilon_t^\top] = 0,\ i > 0$. The Yule-Walker equation for a VAR(1) process is thus given by

$$\Sigma_X^{(i)} = A_1^i \Sigma_X, \tag{52}$$

with $\Sigma_X = Var(X_t)$, which for $i = 1$ implies

$$A_1 = \Sigma_X^{(1)} \Sigma_X^{-1}. \tag{53}$$

To generalize this result to VAR processes of higher orders, the same trick used in the beginning of Section 2 can also be used here to transition from a VAR(p) to a VAR(1) process. Indeed, consider the companion form expressed in (4) where $\Xi_t \sim \mathcal{N}_{kp}(0, \Sigma_\Xi)$ and $\eta_t \sim \mathcal{N}_{kp}(0, \Omega_\Xi)$ with

$$\Sigma_\Xi = \begin{pmatrix} \Sigma_X & \Sigma_X^{(1)} & \cdots & \Sigma_X^{(p-1)} \\ \Sigma_X^{(1)} & \Sigma_X & \cdots & \Sigma_X^{(p-2)} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_X^{(p-1)} & \Sigma_X^{(p-2)} & \cdots & \Sigma_X \end{pmatrix} \text{ and } \Omega_\Xi = \begin{pmatrix} \Omega & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}. \quad (54)$$

Therefore, the Yule-Walker equation for Ξ_t evaluated at $i = 1$ is given by

$$A = \Sigma_\Xi^{(1)} \Sigma_\Xi^{-1}. \quad (55)$$

5.2.2 Sparse Yule-Walker Estimator

To construct an estimator for A, we build on a relation involving A implied by (55), specifically

$$A\Sigma_\Xi - \Sigma_\Xi^{(1)} = 0. \quad (56)$$

Since the population parameters Σ_Ξ and $\Sigma_\Xi^{(1)}$ are presumably not available, we replace them by their sample analog, respectively

$$S \coloneqq (T - p + 1)^{-1} \sum_{t=p}^{T} \xi_t \xi_t^\top \text{ and } S^{(1)} \coloneqq (T - p)^{-1} \sum_{t=p+1}^{T} \xi_t \xi_{t-1}^\top, \quad (57)$$

with (56) becoming

$$AS - S^{(1)} = 0. \quad (58)$$

Since the parameters of interest are the first k rows of A, or $\Phi^\top$, we can focus on first k rows of both sides of (58), or the equation

$$\Phi^\top S - S_{\mathcal{I}}^{(1)} = 0, \quad (59)$$

where $S_{\mathcal{I}}^{(1)}$ corresponds to the submatrix of $S^{(1)}$ indexed by the rows in $\mathcal{I} = \{1, \ldots, k\}$. Then, naturally, we might imagine that a good estimate of $\Phi^\top$ would be one that minimizes a certain norm of the left-hand side matrix in (59), since in the case of the population parameters, this matrix equals zero.

The proposed sparse Yule-Walker estimator (sYWE) of $\Phi^\top$ is

$$\Phi_{\text{sYWE}}^\top = \underset{M \in \mathbb{R}^{k \times kp}}{\arg\min} \ \|M\|_{1,1} \text{ s.t. } \left\| MS - S_{\mathcal{I}}^{(1)} \right\|_{\max} \leq \lambda. \quad (60)$$

Note that this regularized estimation problem is sort of an inverted problem compared to how regularized estimators are usually structured. Indeed, generally, the loss function to be minimized is the main part of the parameter estimation problem and it is chosen such that the resulting estimator satisfies

sought after properties such as unbiasedness or consistency. On the other hand, the regularizer (or the constraint) is what induces the sparsity of the estimator. In the case of the sYWE, the constraint is the main part of the estimation problem whereas the loss function is the inducer of sparsity.

Furthermore, (60) can be equivalently decomposed into sub-problems. Indeed, let $\Phi_{\text{sYWE; i}}^\top$ be the i-th row of $\Phi_{\text{sYWE}}^\top$. Then, each $\Phi_{\text{sYWE; i}}^\top$ can be solved by

$$\Phi_{\text{sYWE; i}}^\top = \underset{B_i \in \mathbb{R}^{kp}}{\arg\min} \ \|B_i\|_1 \ \text{s.t.} \ \left\| B_i S - S_i^{(1)} \right\|_\infty \leq \lambda, \quad i = 1, \dots, k \tag{61}$$

such that $\Phi_{\text{sYWE}}^\top = (\Phi_{\text{sYWE; 1}}, \dots, \Phi_{\text{sYWE; k}})^\top$.

5.2.3 Linear Program

We now show that the constrained minimization problem in (61) can be expressed as a linear program that can be solved efficiently by the well-known Simplex algorithm. Recall that any $a \in \mathbb{R}$ can be written as $a = a^+ - a^-$ where $a^+ = a\mathbb{1}(a > 0)$ and $a^- = -a\mathbb{1}(a < 0)$, so that $|a| = a^+ + a^-$. Then, for any $v = (v_1, \dots, v_d) \in \mathbb{R}^d$, let $v^+ = (v_1^+, \dots, v_d^+)$ and $v^- = (v_1^-, \dots, v_d^-)$. Moreover, for $v, w \in \mathbb{R}^d$, we say that $v \geq 0$ if $v_1, \dots, v_d \geq 0$, $v \leq 0$ if $v_1, \dots, v_d \leq 0$, $v \geq w$ if $v - w \geq 0$ and $v \leq w$ if $v - w \leq 0$. Considering these notations, it follows that (61) can be written as

$$\begin{aligned} \Phi_{\text{sYWE; i}}^\top = \underset{B_i^+ - B_i^- \in \mathbb{R}^{kp}}{\arg\min} \quad & (B_i^+ + B_i^-)\mathbf{1}_{kp} \\ \text{s.t.} \ & \left\| B_i^+ S - B_i^- S - S_i^{(1)} \right\|_\infty \leq \lambda, \ B_i^+, B_i^- \geq 0, \ i = 1, \dots, k, \end{aligned} \tag{62}$$

where $\mathbf{1}_d = (1, \dots, 1)^\top \in \mathbb{R}^d$. Moreover, since the constraint $\|v\|_\infty \leq \lambda$ is equivalent to the constraint imposing that all elements of v be less than λ in absolute value, (62) can be further simplified to

$$\begin{aligned} \Phi_{\text{sYWE; i}}^\top = \underset{B_i^+ - B_i^- \in \mathbb{R}^{kp}}{\arg\min} \quad & (B_i^+ + B_i^-)\mathbf{1}_{kp} \\ \text{s.t.} \quad & (B_i^+ S - B_i^- S - S_i^{(1)})^\top \leq \lambda \mathbf{1}_{kp} \\ & (-B_i^+ S + B_i^- S + S_i^{(1)})^\top \leq \lambda \mathbf{1}_{kp} \\ & B_i^+, B_i^- \geq 0, \ i = 1, \dots, k, \end{aligned}$$

which is equivalent to

$$\Phi_{\text{sYWE; i}} = \underset{b_i^+ - b_i^- \in \mathbb{R}^{kp}}{\arg\min} \ \mathbf{1}_{2kp}^\top \omega \quad \text{s.t.} \ \Lambda + W\omega \geq 0, \tag{63}$$

where $b_i^+ = (B_i^+)^\top$, $b_i^- = (B_i^-)^\top$,

$$\Lambda = \begin{pmatrix} S_i^{(1)\top} + \lambda \mathbf{1}_{kp} \\ -S_i^{(1)\top} + \lambda \mathbf{1}_{kp} \end{pmatrix}, \ W = \begin{pmatrix} -S & S \\ S & -S \end{pmatrix} \text{ and } \omega = \begin{pmatrix} b_i^+ \\ b_i^- \end{pmatrix}.$$

Indeed, the constrained minimization problem in (63) is a linear program which can be solved by the Simplex algorithm.

5.3 PSC-Based sVAR

Another sVAR estimation method which does not rely on neither a multivariate regression model nor on the Yule-Walker equation was introduced in [3]. This method involves the partial spectral coherence (PSC), a function from the frequency-domain of time series, as a statistic for the conditional correlation between two processes. The idea consists of imputing a zero coefficient to pairs not conditionally correlated according to their PSC. In a subsequent step, a refinement based on the significance of estimates is carried out to eliminate remaining potentially spurious non-zero coefficients.

5.3.1 Multivariate Spectral Analysis and PSC

The autocovariance generating function of a k-dimensional process X_t with an absolutely summable sequence autocovariances $\{\Sigma_X^{(l)}\}$, $l \in \mathbb{Z}$, $G_X : \mathbb{C} \to \mathbb{C}^{k\times k}$, is defined as

$$G_X(z) = \sum_{l=-\infty}^{\infty} \Sigma_X^{(l)} z^l, \tag{64}$$

where $\Sigma_X^{(l)} = \mathbb{E}[X_t X_{t-l}^\top]$ and $z \in \mathbb{C}$. Moreover, the population spectrum of X_t, $S_X : \mathbb{R} \to \mathbb{C}^{k\times k}$, is defined as

$$S_X(\omega) = \frac{1}{2\pi} G_X(e^{-i\omega}) = \frac{1}{2\pi} \sum_{l=-\infty}^{\infty} \Sigma_X^{(l)} e^{-i\omega l}, \tag{65}$$

which is the autocovariance generating function evaluated at $z = e^{-i\omega} = \cos\omega - i\sin\omega$, $\omega \in \mathbb{R}$, $i = \sqrt{-1}$, and divided by 2π. It follows, the important result

$$\int_{-\pi}^{\pi} S_X(\omega) e^{i\omega j} d\omega = \Sigma_X^{(l)}, \tag{66}$$

showing that the same information is contained in both the autocovariances and the population spectrum, and implying the special case for $j = 0$

$$\Sigma_X^{(0)} = \int_{-\pi}^{\pi} S_X(\omega) d\omega \tag{67}$$

which effectively shows that $\mathrm{Cov}(X_{i,t}, X_{j,t})$ corresponds to the area under the curve of the (i,j) element of $S_X(\omega)$ on $[-\pi, \pi]$. Thus, the i-th diagonal element of the multivariate spectrum $S_X(\omega)$ corresponds to the univariate spectrum of the scalar process $X_{i,t}$. From the theory of univariate spectral analysis, it follows that every diagonal element of $S_X(\omega)$ is real valued and positive for all ω. This does not necessarily apply to off-diagonal elements of $S_X(\omega)$ which in general

are complex numbers. To that, let us denote the (i,j) element of $S_X(\omega)$ by $S_{X;i,j}(\omega)$, representing the population cross spectrum from $X_{j,t}$ to $X_{i,t}$.

As $X_{i,t}$ and $X_{j,t}$ denote two individual processes of the system X_t, let $X_{-ij,t}$ denote the $(k-2)$-dimensional sub-system of X_t excluding $X_{i,t}$ and $X_{j,t}$. We are interested in computing a measure of the correlation between $X_{i,t}$ and $X_{j,t}$ adjusted for the linear effect of $X_{-ij,t}$ on the latter two. Removing the effect of $X_{-ij,t}$ from $X_{i,t}$ can be done using linear filters. The optimal linear filter for this task is given by

$$\{D_i^{(l)*}\} = \underset{\{D_i^{(l)} \in \mathbb{R}^{k-2}, l \in \mathbb{Z}\}}{\arg\min} \mathbb{E}\left[\left(X_{i,t} - \sum_{l=-\infty}^{\infty} D_i^{(l)} X_{-ij,t-l}\right)^2\right], \tag{68}$$

which in fact represents the set of linear combinations of $X_{-ij,t}$ summed for all t that best explain $X_{i,t}$. The residual process resulting from this optimal linear filter is

$$\varepsilon_{i,t} = X_{i,t} - \sum_{l=-\infty}^{\infty} D_i^{(l)*} X_{-ij,t-l}. \tag{69}$$

The conditional correlation between $X_{i,t}$ and $X_{j,t}$ is simply $\mathrm{Corr}(\varepsilon_{i,t}, \varepsilon_{j,t})$, and we say that $X_{i,t}$ and $X_{j,t}$ are conditionally uncorrelated corrected for $X_{-ij,t}$ if and only if $\mathrm{Corr}(\varepsilon_{i,t}, \varepsilon_{j,t+l}) = 0 \; \forall \; l \in \mathbb{Z}$. This is equivalent to $s_{i,j}(\omega) = 0 \; \forall \; \omega \in [-\pi, \pi]$, where

$$s_{i,j}(\omega) = \frac{1}{2\pi} \sum_{l=-\infty}^{\infty} \gamma_{i,j}(l) e^{-i\omega l} \tag{70}$$

denotes the population cross spectrum from $\varepsilon_{j,t}$ to $\varepsilon_{i,t}$, with

$$\gamma_{i,j}(l) = \mathrm{Cov}(\varepsilon_{i,t}, \varepsilon_{j,t-l}).$$

The PSC between $X_{i,t}$ and $X_{j,t}$ is defined as the following scaled cross spectrum between $\varepsilon_{i,t}$ and $\varepsilon_{j,t}$

$$\mathrm{PSC}_{i,j}(\omega) = \frac{s_{i,j}(\omega)}{\sqrt{s_{i,i}(\omega) s_{j,j}(\omega)}} \tag{71}$$

and can be rewritten as

$$PSC_{i,j}(\omega) = -\frac{R_{X;i,j}(\omega)}{\sqrt{R_{X;i,i}(\omega) R_{X;j,j}(\omega)}}, \tag{72}$$

where $R_{X;i,j}$ denotes the (i,j) element of S_X^{-1}. The formulation in (72) is computationally more convenient since it requires only the computation of the system's spectrum and its inversion once for all possible (i,j) pairs, whereas the computation of (71) would require finding different residual cross spectrums for each different pair. Obviously, $X_{i,t}$ and $X_{j,t}$ are conditionally uncorrelated if and only if $R_{X;i,j}(\omega) = 0 \; \forall \; \omega \in [-\pi, \pi]$.

5.3.2 Sparse PSC Estimator

Step 1 - PSC Filtering

We follow the rule

$$A_{i,j}^{(l)} = A_{j,i}^{(l)} = 0,\ i \neq j,\ l = 1, \ldots, p \quad \text{if } \mathrm{PSC}_{i,j}(\omega) = \mathrm{PSC}_{j,i}(\omega) = 0,\ \forall\ \omega \in [-\pi, \pi], \tag{73}$$

where $A_{i,j}^{(l)}$ denotes the (i,j) element of A_l. Effectively, if $A_{i,j}^{(l)}$ is set to zero, so is $A_{j,i}^{(l)}$, since $R_{X;i,j}$ equals $R_{X;j,i}$, meaning that zero coefficients in the transition matrices are set symmetrically. Also, since $R_{X;i,j}(\omega)$ being constant at zero over all frequencies ω implies that $X_{i,t}$ and $X_{j,t}$ are conditionally uncorrelated at all lags, we set the (i,j) and (j,i) coefficients to zero in all transition matrices.

However, the PSC being a function of the population spectrum of X_t solely, the latter must be estimated since it depends on the unavailable population parameters $\Sigma_{X;l}$. We estimate $S_X(\omega)$ as

$$\widehat{S}_X(\omega) = \frac{1}{2\pi}\left[\widehat{\Sigma}_X^{(0)} + \sum_{l=1}^{T-1} k_l \left(\widehat{\Sigma}_X^{(l)} e^{-i\omega l} + \widehat{\Sigma}_X^{(l)\top} e^{-i\omega l}\right)\right], \tag{74}$$

where $\widehat{\Sigma}_X^{(l)} = (T-l)^{-1}\sum_{t=l+1}^{T} X_t X_{t-l}^\top$ and $k_l = \sin(\nu l)/\nu l$, $\nu \in \mathbb{R}$ is a smoothing kernel. The estimated PSC becomes

$$\widehat{\mathrm{PSC}}_{i,j}(\omega) = -\frac{\widehat{R}_{X;i,j}(\omega)}{\sqrt{\widehat{R}_{X;i,i}(\omega)\widehat{R}_{X;j,j}(\omega)}}, \tag{75}$$

where $\widehat{R}_X = \widehat{S}_X^{-1}$. Of course, due to sampling variability, the estimate of the PSC of a given pair might not be zero over all frequencies even if the pair is in fact conditionally uncorrelated. Therefore, deprived of distributional knowledge that would allow us to conjecture the significance of an estimated PSC, we simply assume that the farther the estimate is from zero, the less likely is the pair to be conditionally uncorrelated. Though, since the PSC is a function of the frequency ω, a statistic quantifying its deviance from zero over multiple frequencies is required. The proposed statistic is

$$\widehat{s}_{i,j} = \max_x |\widehat{\mathrm{PSC}}_{i,j}(x)|^2, \tag{76}$$

where $x \in \{2\pi l/T : \ l = 1, \ldots, T\}$. Then, a sequence $\mathcal{Q}_1$ containing the $\binom{k}{2} = k(k-1)/2$ distinct pairs is constructed by ordering their associated $\widehat{s}_{i,j}$ increasingly. As a rule, we impute zero coefficients to the m first elements of $\mathcal{Q}_1$, with m effectively being a tuning parameter. The other $k^2p - 2pm$ non-zero coefficients are estimated by a constrained MLE.

Consider the VAR(p) model represented as in (3) where zero and non-zero coefficients are chosen by the above rule. Then, denote $\phi = \mathrm{vec}_{k\times kp}(\Phi^\top)$, which

contains $n = k^2p - 2pm$ freely varying parameters $A_{i,j}^{(l)}$ and $2pm$ parameters fixed to zero. The goal is to compute the maximum likelihood estimate of the non-zero coefficients subject to the constraint that the other coefficients must be zero. For this, we rewrite ϕ as

$$\phi = R\gamma, \tag{77}$$

where $R \in \{0,1\}^{k^2p \times n}$ is the constraint matrix and $\gamma \in \mathbb{R}^n$ contains only the freely varying coefficients in ϕ. Each row of R contains at most one 1 and at a different column, thus R is of full column rank. As an illustration, consider a simple 2-dimensional sVAR(1) characterized by

$$X_t = \begin{pmatrix} A_{1,1}^{(1)} & 0 \\ A_{2,1}^{(1)} & A_{2,2}^{(1)} \end{pmatrix} X_{t-1} + \epsilon_t. \tag{78}$$

In this case, $k^2p = 4$, $m = 1$ and $n = 3$, and $\phi = R\gamma$ would be given by

$$\begin{pmatrix} A_{1,1}^{(1)} \\ 0 \\ A_{2,1}^{(1)} \\ A_{2,2}^{(1)} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} A_{1,1}^{(1)} \\ A_{2,1}^{(1)} \\ A_{2,2}^{(1)} \end{pmatrix}. \tag{79}$$

It turns out that, under the constraint in (77), the MLE of ϕ and that of $\Omega = \mathrm{Var}(\epsilon_t)$ are given by (see [10] p. 200)

$$\widehat{\phi} = R\left(R^\top (X^\top X \otimes \widehat{\Omega}^{-1}) R\right)^{-1} R^\top (X^\top \otimes \widehat{\Omega}^{-1}) \operatorname{vec}(Y^\top) \tag{80}$$

$$\text{and} \quad \widehat{\Omega} = (T-p)^{-1} \sum_{t=p+1}^{T} (X_t - \widehat{X}_t)(X_t - \widehat{X}_t)^\top, \tag{81}$$

where $\widehat{X}_t = \widehat{\Phi}_{\mathrm{PSC}}^\top \Xi_{t-1}$ with $\widehat{\Phi}_{\mathrm{PSC}}^\top = \mathrm{vec}_{k \times kp}^{-1}(\widehat{\phi})$, and X as well as Y are as given in (14). Since (80) appears in (81) and vice versa, we must iteratively update both with the other's last value until either changes are small enough.

Step 2 - refinement

Since the PSC can not be applied to pairs (i,j) when $i = j$ and the first step might leave spurious non-zero coefficients, we proceed to a second step to eliminate those coefficients among the $k^2p - 2pm$ remaining non-zero coefficients $A_{i,j}^{(l)}$. In this step, a sequence is once again constructed, this time using a statistic measuring the significance of the non-zero parameters estimated in step 1. The statistic used is the absolute value of the t-statistic given by

$$|t_{i,j}^{(l)}| = \left| \frac{\widehat{A}_{i,j}^{(l)}}{\mathrm{SE}(\widehat{A}_{i,j}^{(l)})} \right|, \tag{82}$$

where $\mathrm{SE}(\widehat{A}_{i,j}^{(l)})$ is approximated from the asymptotic distribution of the constrained MLE of step 1 (see [10] p. 201)

$$\sqrt{T}\left(\widehat{\phi}-\phi\right) \xrightarrow{d} \mathcal{N}_{k^2p}\left(0, R\left(R^\top(\Sigma_{\Xi;0}\otimes\Omega^{-1})R\right)^{-1}R^\top\right) \tag{83}$$

by replacing $\Sigma_\Xi^{(0)}$ and Ω by their MLE. The sequence $\mathcal{Q}_2$ is created by ordering the triplets (l,i,j) increasingly according to their t-statistic in (83). Then, the first q triplets are imputed a zero coefficient as their coefficient are considered to be the less significant, with q thus being a tuning parameter.

6 Indirect Cointegration Estimation

As seen so far, all cointegration estimation methods based on Johansen's reduced VECM representation, sparse or non-sparse, are used to estimate the cointegrating matrix β directly. In other words, the cointegrating matrix appears explicitly in the estimation problem and an optimization problem is solved directly with respect to the cointegrating matrix, leading to a direct estimate of it. Indeed, the non-sparse Johansen's DiCE bases itself on the maximization of the likelihood of a reduced rank VECM, in which the long term effect matrix Π is already decomposed as $\alpha\beta^\top$. Then, as we have seen, the β maximizing this likelihood are the eigenvectors solution to a generalized eigenvalue problem. In the sparse setting, Wilms' DiCE maximizes a constrained penalized likelihood in which the VECM is unrestricted and an L^1 penalty is applied to β, under the constraint $\Pi = \alpha\beta^\top$. This has the effect that some elements of the cointegrating vectors will be estimated as zero.

In this thesis, we introduce a new technique to estimate the cointegrating matrix indirectly. The indirect cointegration estimation (ICE) consists of first estimating a VAR model and then rewriting it as VECM with the estimated VAR transition matrices. This gives an estimate for Π, which due to sampling variability, will not necessarily be of the rank of interest $r < k$. For this reason, we proceed to a low-rank approximation of Π, for a chosen rank r, $\widetilde{\Pi}$, such that $\operatorname{rank}(\widetilde{\Pi}) = r$. Then, to obtain the cointegrating matrix, $\widetilde{\Pi}$ is factorized as $AB^\top$ so that B corresponds to our indirect estimate of the cointegrating matrix.

The ICE technique can be applied to both the sparse or the non-sparse setting. In the non-sparse setting, this technique would be an indirect analogue to Johansen's DiCE. In the sparse setting though, the ICE is not an analogue of Wilms' DiCE since there is no guarantee that the ICE cointegrating vectors contain zero coefficients. Indeed, the inherent hypothesis of the ICE is that the underlying VAR structure is sparse, meaning that the transition matrices contain zero coefficients. This does not imply that zero coefficients are present in the cointegrating matrix.

6.1 Low-Rank Approximation

Consider the VAR(p) in (1) and the estimated transition matrices $\widehat{A}_1, \ldots, \widehat{A}_p$, from any estimator. The estimated VECM (as in (25)) resulting from this VAR estimation is characterized by the following parameters

$$\widehat{\Pi} = -I_k + \sum_{i=1}^{p} \widehat{A}_i, \quad \widehat{\Gamma}_i = -\sum_{j=i+1}^{p} \widehat{A}_j. \tag{84}$$

We wish to find a matrix $\widetilde{\Pi}_r \in \mathbb{R}^{k\times k}$ of rank r which is the closest to $\widehat{\Pi}$ in the sense of some matrix norm. One way to formulate this problem can be as follows

$$\widetilde{\Pi}_r = \underset{D\in\mathbb{R}^{k\times k}}{\arg\min} \left\| \widehat{\Pi} - D \right\|_{\mathrm{F}} \text{ s.t. } \operatorname{rank}(D) = r. \tag{85}$$

The Frobenius norm is convenient in this case since this constrained optimization has an explicit solution, provided by the Eckart-Young-Mirsky theorem. The theorem tells us that the solution of (85) stems directly from the singular value decomposition of the matrix we want to approximate, $\widehat{\Pi}$. Indeed, let

$$\widehat{\Pi} = U\Sigma V^\top \tag{86}$$

be the singular value decomposition of $\widehat{\Pi}$. We know that $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_k)$ with $\sigma_1 > \ldots > \sigma_k$ being the singular values of $\widehat{\Pi}$. Also, the number of non-zero singular values corresponds to the rank of $\widehat{\Pi}$. The theorem states that

$$\widetilde{\Pi}_r = U\widetilde{\Sigma}V^\top, \tag{87}$$

where

$$\widetilde{\Sigma} = \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} \tag{88}$$

with $\Sigma_r = \text{diag}(\sigma_1, \ldots, \sigma_r)$, is such that

$$\left\| \widehat{\Pi} - \widetilde{\Pi}_r \right\|_{\text{F}} = \min_{\text{rank}(D)=r} \left\| \widehat{\Pi} - D \right\|_{\text{F}} = \sqrt{\sum_{i=r+1}^{k} \sigma_i} \,. \tag{89}$$

6.2 Rank Factorization

Now that we have an approximation of $\widehat{\Pi}$ which is of rank r, we can easily decompose it as the reduced VECM form of Johansen requires to retrieve the cointegrating matrix. Indeed, $\text{rank}(\widetilde{\Pi}_r) = r$ so that it permits the factorization of the form $\widetilde{\Pi}_r = \alpha\beta^\top$, where $\alpha, \beta \in R^{k\times r}$.

To do so, let F be the reduced row echelon form of $\widetilde{\Pi}_r$. Then, A is the sub-matrix of $\widetilde{\Pi}_r$ that does not contain the columns which are not pivot columns of F, and $\beta^\top$ is the sub-matrix of F that does not contain the rows comprised of only zeros. Therefore, β is the cointegrating matrix stemming from the ICE.

7 Simulation Study

The goal of this simulation study is to compare the performance of the various presented estimators with respect to their respective cointegration estimates, across four different data generating processes (DGPs). To do this, the true cointegrating matrices are needed in order to be able to compare them with the estimated ones. This motivates the following simulation set up.

The four different DGPs are all cointegrated VAR(1) processes but differ in dimensionality and sparsity. The low dimensional setting consists of $k = 50$ variables and $T = 200$ observations, whereas the high dimensional one has the same number of variables but $T = 40$ observations. Sparsity in the sparse settings is induced through the cointegrating matrices, which as we will see, has an impact on the sparsity structure of the transition matrices. Each of the four DGPs are simulated $M = 100$ times and the cointegrating rank is set to $r = 1$.

Since the simulation method is invariant with respect to dimensionality, the following sub-section describes how the simulations are performed in the non-sparse and sparse case.

7.1 Transition Matrix Simulation

7.1.1 Non-Sparse Setting

Since $p = 1$, $A = A_1 \in \mathbb{R}^{k \times k}$. For each iteration $m \in \{1, \dots, M\}$, A is simulated using the following algorithm.

Algorithm 1: Generate non-explosive transition matrix

$A \leftarrow 0_{k,k}$
$expl \leftarrow$ TRUE
while $expl$ **do**
  for each i in $\{1, \dots, k\}$ **do**
    $a_{i,i} \leftarrow$ uniform on $[0.8, 1]$
    $left \leftarrow 1 - a_{i,i}$
    for each j in $\{1, \dots, k\} \setminus \{i\}$ **do**
      $a_{i,j} \leftarrow$ uniform on $[-2 \cdot left/(k-1), 2 \cdot left/(k-1)]$
    end for
  end for
  $eigs \leftarrow$ Moduli of eigenvalues of A
  if $\max(eigs) < 1$ **then**
    $expl \leftarrow$ FALSE
  end if
end while

The algorithm iterates in order respectively over the rows and the columns of the matrix, while for the columns, it first fills the diagonal element of that row with a larger value than other elements of the row, which are then filled with smaller values depending on the value the diagonal element was attributed.

The reason for this is to generate processes that are close to the unit root, thus avoiding trivial stationary systems. The algorithm repeats until it generates a non-explosive transition matrix.

We then retrieve the cointegrating vector ($r = 1$) β and the resulting transition matrix $\tilde{A} = \tilde{\Pi}_r + I_k$ using the rank factorization method presented in Section 6. We consider β to be the true cointegrating vector to which we will compare the estimated cointegrated vector. The coefficient matrix used to simulate the VAR(1) sample is $\tilde{A}$ (not A).

It is important to note that although A is non-explosive, $\tilde{A}$ is. Indeed, as $\text{rank}(\tilde{\Pi}_1) = 1$, $\tilde{\Pi}_1$ has $k - 1$ zero eigenvalues. Since I_k and $\tilde{\Pi}_1$ are commuting matrices and they are summed to get $\tilde{A}$, the latter's eigenvalues are sums of eigenvalues of the summands, and therefore the largest eigenvalue of the sum is less or equal to the sum of the largest eigenvalues of the summands. Empirically, we have observed that in the great majority of cases $\tilde{A}$ has $k - 1$ eigenvalues equal to one and the remaining less than one.

7.1.2 Sparse Setting

We proceed as in 7.1.1 to simulate a non-explosive transition matrix and as in Section 6 to retrieve α and β from that generated matrix. To induce sparsity in the cointegrating vector β, we apply the following algorithm.

Algorithm 2: Transition matrix under sparse cointegration

```
Apply algorithm (1) and rank factorization to retrieve α and β
B̃ ← β
for each i in {1, ..., k} do
    s ← Bernoulli with probability 1/2
    if s is 0 then
        b̃_i ← 0
    end if
end for
Ã ← αB̃^⊤ + I_k
```

Thus, to generate transition matrices for the sparse case, we start by generating transition matrices as in the non-sparse case and we apply rank factorization to retrieve α and β. To induce sparsity in the cointegration, we chose at random components of β and change their value to zero with probability $1/2$. We then simply reconstruct the transition matrix $\tilde{A}$ that will be used for simulations in the sparse case with the new ground truth sparse cointegrating vector $\tilde{B}$.

Note that since $\alpha, \tilde{B} \in \mathbb{R}^k$ and $\tilde{B}$ contains zero entries, the columns of $\tilde{A}$ corresponding to zero entries in $\tilde{B}$ will contain only zeros except for the diagonal entry on that column which will be a one, since the identity matrix is added. Intuitively, this structure implies that the variables that do not enter the cointegrating relationship are independent random walks. Since the probability that a component of the cointegrating vector is set to zero is $1/2$, half of the

processes in the sparse systems will be such random walks on average, and will not influence other cointegrated processes whose corresponding cointegrating coefficient is not zero.

To better visualize the transition matrix in both sparsity settings, consider a grid (see figure 1) with each small square representing a coefficient (component) of the transition matrix. The coloring of the squares has been simplified to highlight the general structure of the matrix. For a given coefficient A_{ij}, it is colored black if $|A_{ij}| \in [0.8, 1]$, gray if $|A_{ij}| \in (0, 0.8)$ and white if $A_{ij} = 0$. As we can see, black coefficient all lie on the diagonal in both cases as it is implied by the generating algorithms 1 and 2. In the sparse setting, if one square is white, all the other squares in that column are also white, as explained in the preceding paragraph.

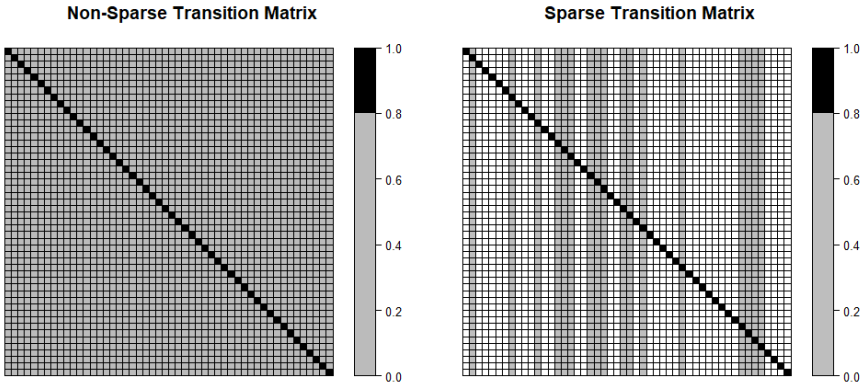

Figure 1: Example of transition matrices in both sparsity settings

7.1.3 VAR(1) Processes Simulation

A total of 400 VAR(1) processes (100 for each DGP set-up) are simulated using multivariate white noise as residuals, i.e. $\epsilon_t \in \mathbb{R}^k \sim \mathcal{N}_k(0, \Omega)$ with $\Omega = I_k$.

7.2 Fitting Procedures and Tuning Parameters

7.2.1 Estimators Fitted

Now that we have generated processes each with an underlying ground truth cointegrating vector, the goal is now to fit the various estimators to these sampled processes to get estimated values for that process' cointegrating vector. Below are the estimators that we apply in each set-up.

		DiCE		ICE		
		Joha	sDiCE	Lasso	sYWE	sPSC
Low	Non-sparse	✓	✓	✓	✓	✓
	Sparse	✓	✓	✓	✓	✓
High	Non-sparse	✗	✓	✓	✓	✓
	Sparse	✗	✓	✓	✓	✓

Table 1: Types of estimators applied for each DGP

7.2.2 Choosing Tuning Parameters

The regularized estimators that we use for the ICE are functions of tuning parameters that vary freely and which we need to fix. These parameters dictate the extent to which the estimator is regularized, and thus, in our case, increase or decrease the sparsity of the estimate. Since in real applications the ground truth cointegrating vector is rarely known, we take great care in this simulation not to optimize the loss, which depends on the ground truth as well as the estimate, with respect to the tuning parameter. Instead, we only use what is observable, namely the generated processes, to tune these parameters with the hope that observed process conveys information on the structure of ground truth cointegration relationship and therefore choose the value of the tuning parameter that best fits the observed process.

To that end, let Λ be the set of values that a tuning parameter λ can take value in. For each estimator and process, we get an estimate of the transition matrix $\hat{A}(\lambda)$ which depends on the tuning parameter. We then build a forecast process $\hat{X} = (\hat{X}_1, \dots, \hat{X}_T)^\top \in \mathbb{R}^{T \times k}$ where

$$\begin{aligned} \hat{X}_2 &= \hat{A}(\lambda) X_1 \\ \hat{X}_t &= \hat{A}(\lambda) \hat{X}_{t-1}, \ \forall\, t > 2 \end{aligned}$$

with X_1 being the first observation of the observed process. We then choose the optimal tuning parameter as

$$\lambda^* = \underset{\lambda \in \Lambda}{\arg\min} \left\| X - \hat{X} \right\|_{\mathrm{F}}. \tag{90}$$

7.3 Performance Measure

Then, for each process and each estimator, we get an estimate of the cointegrating vector $\hat{B}$ which we wish to compare to the ground truth $\tilde{B}$. To assess the performance of an estimator, we use the log of the L^1-norm of the difference between the ground truth and the estimate as the loss function, i.e.

$$l_B = \log \left\| \tilde{B} - \hat{B} \right\|_1 = \log \sum_{i=0}^{k} |\tilde{b}_i - \hat{b}_i|. \tag{91}$$

Naturally, the lower l_B is, the higher the performance. Thus, for each estimator, we get 400 of such losses (except for Johansen's estimator for which only 200 are computable).

8 Results

We find that the sparse direct cointegration estimator (sDiCE) of [15] (highlighted in red in the figures below) seems to perform best in all settings in terms of the median l_B. In the low-dimensional setting, Johansen, Lasso and sPSC all come after sDiCE and have a very similar performance, except in the sparse case where sPSC's performance gets worse relative to Johansen and Lasso. In the high-dimensional setting, Lasso and sPSC are very similar both in the non-sparse and sparse case. In all settings, it is the sYWE that performs worst. This could be due to the strong stationarity (or non-explosivity) assumption implied by the Yule-Walker equation, which as mentioned in section 7.1.1, is not satisfied for the simulated cointegrated processes.

It is important to note though, that these differences in performance might not be significant considering how slight they are and the large variance of all the estimators. This indicates that the indirect cointegration estimation method (ICE) works at least similarly to other already established direct cointegration estimation methods (DiCE).

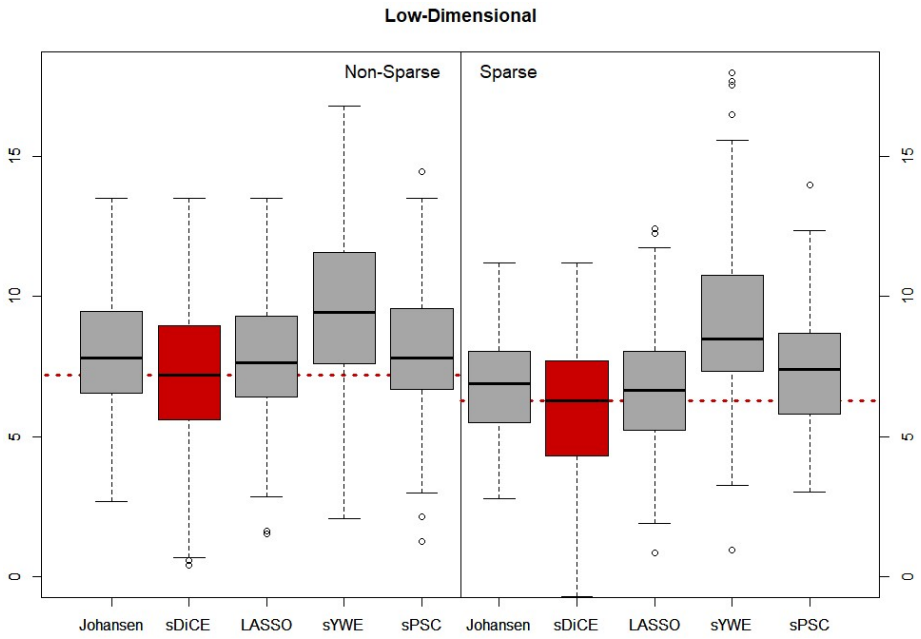

Figure 2: Box plot of l_B of estimators in the low-dimensional setting

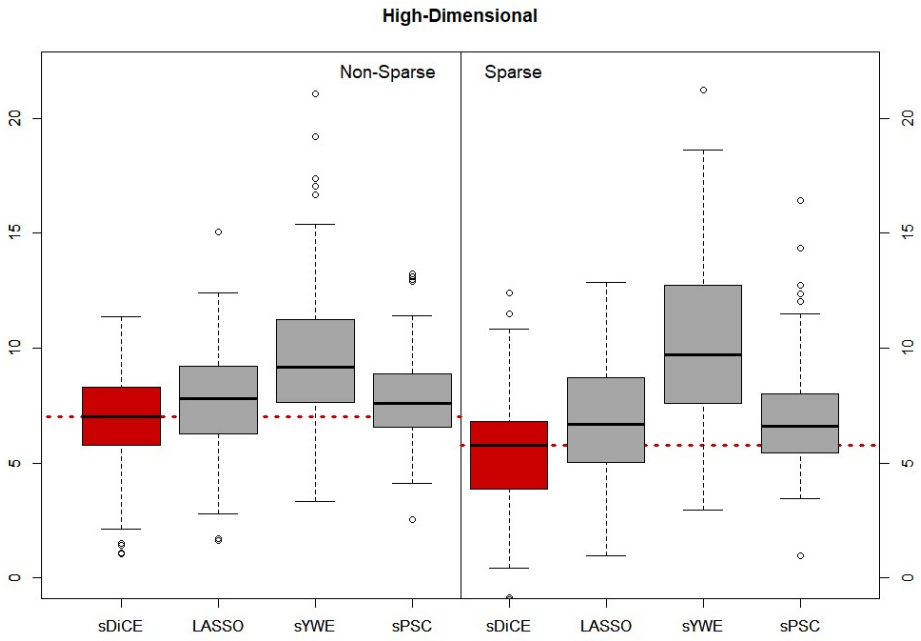

Figure 3: Box plot of l_B of estimators in the high-dimensional setting

Below is one example of a cointegrated VAR(1) system drawn from our simulations in the low-dimensional and non-sparse setting.

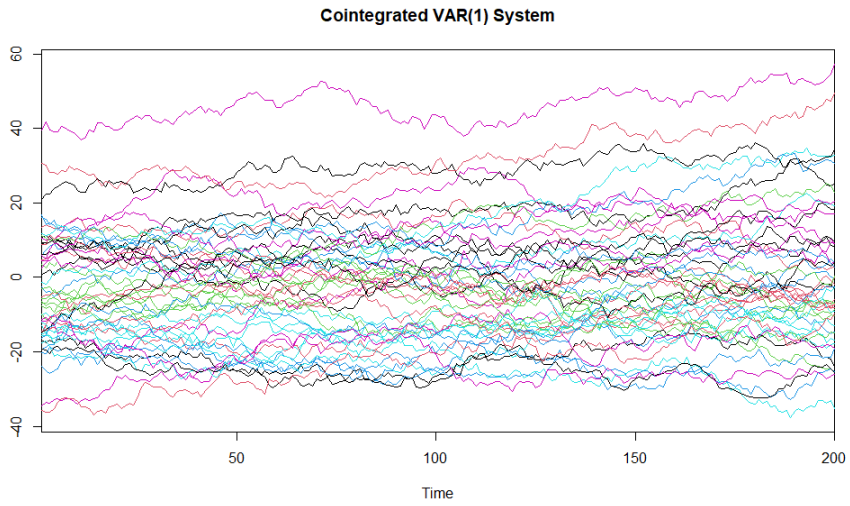

Figure 4: Simulated cointegrated VAR(1) of 50 variables

We also show the resulting cointegrating processes associated to this system. These are in-sample processes, meaning they have been constructed using the same system (see figure 4) on which the cointegrating vectors were estimated. On top of the estimated cointegrating processes, we also plot the ground truth cointegrating process, i.e. using the cointegrating vector implicit the above cointegrated VAR(1) (see section 7.1.1).

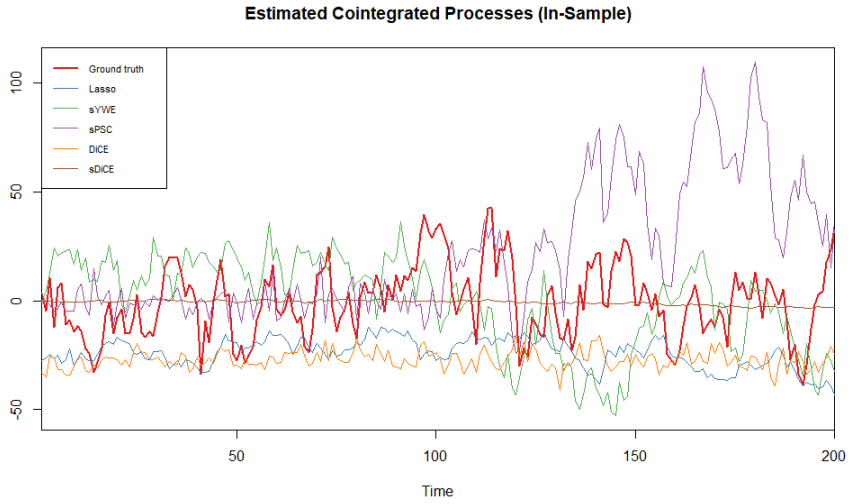

Figure 5: Estimated cointegrating process from each estimator

As we can see, the estimators yield quite different processes compared to the ground truth cointegrating process, although for the most part they are not explosive. Even the benchmark DiCE method fails to adequately fit the true cointegrating process, exhibiting a comparable process as Lasso. The best performing method in terms of l_B, sDiCE, strangely produces a very low variance cointegrating process, suggesting that there might exist a better performance measure to assess one estimator's performance. It is surprising to see that the estimator that seems to track best the true cointegrating process is sYWE, the worst performer in terms of l_B.

9 Conclusion

This thesis shows that circumventing the benchmark direct cointegration estimation method through the reduced form VECM by taking advantage of the equivalence between the latter and a standard VAR yields similar results in terms of performance, as might be expected given the mentioned equivalency. Indeed, rather than estimating directly the cointegrating matrix, we showed that it is possible to estimate the coefficients of a VAR in a first step before applying operations to retrieve the cointegrating matrix. Thanks to this greater flexibility, it might be possible to further increase the performance of cointegration estimations by optimizing the estimation of the system itself, which cannot be done with direct cointegration estimation techniques. The choice of first step estimators in the ICE can be a topic for further research. In addition, this thesis demonstrates the great difficulty that existing cointegration estimation methods have in accurately estimating the cointegrating relationships of reasonably large cointegrated systems. Indeed, this is highlighted by the simulation study in which truly cointegrated systems were simulated in different dimensional and sparse settings. Although both DiCE and ICE yield similar results in terms of the total estimation error of cointegrating vectors, the resulting cointegrating processes differ greatly, which can be problematic in applications where equilibrium relationships are needed or in signal processing. In light of that shortcoming, drastically new methods for cointegration estimation is also an important area of research.

References

[1] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley, 2003.

[2] Philip J Brown, James V Zidek, et al. "Adaptive multivariate ridge regression". In: *The Annals of Statistics* 8.1 (1980), pp. 64–74.

[3] Richard A Davis et al. "Sparse vector autoregressive modeling". In: *Journal of Computational and Graphical Statistics* 25.4 (2016), pp. 1077–1096.

[4] Bradley Efron et al. "Least angle regression". In: *The Annals of statistics* 32.2 (2004), pp. 407–499.

[5] Robert F Engle and Clive WJ Granger. "Co-integration and error correction: representation, estimation, and testing". In: *Econometrica: journal of the Econometric Society* (1987), pp. 251–276.

[6] W.A. Fuller. *Introduction to Statistical Time Series*. Wiley Series in Probability and Statistics. Wiley, 1995.

[7] James Douglas Hamilton. *Time Series Analysis*. Princeton university press, 2020.

[8] Fang Han et al. "A direct estimation of high dimensional stationary vector autoregressions". In: *Journal of Machine Learning Research* (2015).

[9] Søren Johansen et al. *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford University Press on Demand, 1995.

[10] Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media, 2005.

[11] K.V. Mardia et al. *Multivariate Analysis*. Probability and Mathematical Statistics : a series of monographs and textbooks. Academic Press, 1979.

[12] Hsu Nan-Jung et al. "Subset selection for vector autoregressive processes using lasso". In: *Computational Statistics & Data Analysis* 52.7 (2008), pp. 3645–3657.

[13] Michael Osterwald-Lenum et al. "A note with quantiles of the asymptotic distribution of the maximum likelihood cointegration rank test statistics". In: *Oxford Bulletin of Economics and statistics* 54.3 (1992), pp. 461–472.

[14] Y. Sun et al. "Regularization in High-dimensional Statistics". PhD thesis. 2015.

[15] Ines Wilms and Christophe Croux. "Forecasting using sparse cointegration". In: *International Journal of Forecasting* 32.4 (2016), pp. 1256–1267.

A 1

We show here that $(\Phi^*) = \widehat{\Phi}$ where

$$(\Phi^*)^\top = \left(\left(X^\top X\right)^{-1} X^\top Y\right)^\top = Y^\top X \left(X^\top X\right)^{-1} \quad \text{and}$$
$$\widehat{\Phi}^\top = \left(\sum_{t=1}^{T} x_t \xi_{t-1}^\top\right)\left(\sum_{t=1}^{T} \xi_{t-1}\xi_{t-1}^\top\right)^{-1}.$$

Indeed,

$$Y = \underset{T\times k}{\begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_T^\top \end{pmatrix}} \quad \text{and} \quad X = \underset{T\times kp}{\begin{pmatrix} \xi_0^\top \\ \xi_1^\top \\ \vdots \\ \xi_{T-1}^\top \end{pmatrix}},$$

such that

$$Y^\top X = \begin{pmatrix} x_1 & x_2 & \cdots & x_T \end{pmatrix} \begin{pmatrix} \xi_0^\top \\ \xi_1^\top \\ \vdots \\ \xi_{T-1}^\top \end{pmatrix} = \sum_{t=1}^{T} x_t \xi_{t-1}^\top \qquad \text{and}$$

$$X^\top X = \begin{pmatrix} \xi_0 & \xi_1 & \cdots & \xi_{T-1} \end{pmatrix} \begin{pmatrix} \xi_0^\top \\ \xi_1^\top \\ \vdots \\ \xi_{T-1}^\top \end{pmatrix} = \sum_{t=1}^{T} \xi_{t-1}\xi_{t-1}^\top$$

■

B 2

We show here that $\widehat{\Omega} = \Omega^*$ where

$$\widehat{\Omega} = T^{-1}\sum_{t=1}^{T} \widehat{\epsilon}_t \widehat{\epsilon}_t^\top, \quad \Omega^* = T^{-1} Y^\top P Y \tag{92}$$

with

$$\widehat{\epsilon}_t = x_t - \widehat{\Phi}^\top \xi_{t-1}, \quad P = I_T - X\left(X^\top X\right)^{-1} X^\top,$$
$$Y = \begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_T^\top \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} \xi_0^\top \\ \xi_1^\top \\ \vdots \\ \xi_{T-1}^\top \end{pmatrix}.$$

Indeed,

$$\Omega^* = T^{-1}Y^\top PY = T^{-1}\left[Y^\top Y - Y^\top X(X^\top X)^{-1}X^\top Y\right]$$

and

$$\begin{aligned}
\widehat{\Omega} &= T^{-1}\sum_{t=1}^{T}\widehat{\epsilon}_t\widehat{\epsilon}_t^{\top} \\
&= T^{-1}\left[\sum_{t=1}^{T}\left(x_t - \widehat{\Phi}^\top\xi_{t-1}\right)\left(x_t^\top - \xi_{t-1}^\top\widehat{\Phi}\right)\right] \\
&= T^{-1}\left[\sum_{t=1}^{T}x_t x_t^\top - \left(\sum_{t=1}^{T}x_t\xi t - 1^\top\right)\widehat{\Phi} - \widehat{\Phi}^\top\sum_{t=1}^{T}\xi_{t-1}x_t^\top \right. \\
&\qquad \left. + \widehat{\Phi}^\top\left(\sum_{t=1}^{T}\xi_{t-1}\xi_{t-1}^\top\right)\widehat{\Phi}\right] \\
&= T^{-1}\left[Y^\top Y - Y^\top X\widehat{\Phi} - \widehat{\Phi}^\top X^\top Y + \widehat{\Phi}^\top X^\top X\widehat{\Phi}\right] \\
&= T^{-1}\left[Y^\top Y - Y^\top X\left(X^\top X\right)^{-1}X^\top Y - Y^\top X\left(X^\top X\right)^{-1}X^\top Y \right. \\
&\qquad \left. + Y^\top X\left(X^\top X\right)^{-1}X^\top X\left(X^\top X\right)^{-1}X^\top Y\right] \\
&= T^{-1}\left[Y^\top Y - Y^\top X(X^\top X)^{-1}X^\top Y\right]
\end{aligned}$$

■