# Convex Estimation of Cointegrated VAR Models by a Nuclear Norm Penalty

**M. Signoretto** * **and J. A. K. Suykens** *

* *Katholieke Universiteit Leuven, ESAT-SCD/SISTA*
*Kasteelpark Arenberg 10, B-3001 Leuven (BELGIUM)*
*Email: marco.signoretto@esat.kuleuven.be and*
*johan.suykens@esat.kuleuven.be*

**Abstract:** Cointegrated Vector AutoRegressive (VAR) processes arise in the study of long run equilibrium relations of stochastic dynamical systems. In this paper we introduce a novel convex approach for the analysis of these type of processes. The idea relies on an error correction representation and amounts at solving a penalized empirical risk minimization problem. The latter finds a model from data by minimizing a trade-off between a quadratic error function and a nuclear norm penalty used as a proxy for the cointegrating rank. We elaborate on properties of the proposed convex program; we then propose an easily implementable and provably convergent algorithm based on FISTA. This algorithm can be conveniently used for computing the regularization path, i.e., the entire set of solutions associated to the trade-off parameter. We show how such path can be used to build an estimator for the cointegrating rank and illustrate the proposed ideas with experiments.

## 1. INTRODUCTION

Unit root nonstationary multivariate processes play an important role in the study of dynamical stochastic systems Box and Tiao [1977], Engle and Granger [1987], Stock and Watson [1988], Johansen [1988]. Contrary to their stationary counterpart these processes are allowed to have trends or shifts in the mean or in the covariances. This feature makes them suitable to describe many phenomena of interest such as economic cycles and population dynamics. In this paper we focus on VAR processes. It is well known that these processes can generate stochastic and deterministic trends if the associated polynomial matrix has zeros on the unit circle. If some of the variables within the same VAR process move together in the long-run, they are called *cointegrated*. This situation is of considerable practical interest. Equilibrium relationships arise between economic variables such as, for instance, household income and expenditures. Cointegration has also been advocated to describe long-term parallel growth of mutually dependent indicators such as regional population and employment growth or city populations and total urban populations Payne and Ewing [1997], Sharma [2003], Møller and Sharp [2008].

⋆ Research supported by the Research Council KUL: GOA /11/05 Ambiorics, GOA/10/09 MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC) en PFV/10/002 (OPTEC). Flemish Government: FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel). Research communities (WOG: ICCoS, ANMMM, MLDM). Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011); IBBT EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940), FP7-SADCO ( MC ITN-264735), ERC HIGHWIND (259 166). The scientific responsibility is assumed by its authors.

The analysis of cointegrated VAR models present challenges that are not present in the stationary case. In particular one of the main goal in the analysis of cointegrated system is the estimation of the cointegrating rank. In this work we propose a novel approach that relies on an error correction representation of cointegrated VAR processes. The approach consists of solving a convex program and uses a nuclear norm as a proxy for the cointegrating rank. We show how the regularization path arising from different values of a trade-off parameter can be used to estimate the cointegrating rank. In order to compute solutions we propose to use a simple yet efficient algorithm based on an existing procedure called FISTA (*fast iterative shrinkage-thresholding algorithm*).

In Section 2 we recall the concept of cointegrated VAR models, error correction representations and cointegrating rank. In Section 3 we present our main problem formulation and discuss its properties. Section 4 deals with an algorithm to compute solutions. In Section 5 we introduce an estimator for the cointegrating rank based on the regularization path. In Section 6 we report experiments. We conclude with final remarks in Section 7.

## 2. COINTEGRATED VAR($P$) MODELS

In the following we denote vectors as lower case letters $(a, b, c, \ldots)$ and matrices as capital letters $(A, B, C, \ldots)$. In particular, we use $I$ to indicate the identity matrix, the dimension of which will be clear from the context. For a positive integer $P$ we write $\mathbb{N}_P$ to denote the set $\{1, \ldots, P\}$. Finally we use $\langle A, B \rangle$ to denote the inner product between $A, B \in \mathbb{R}^{D_1 \times D_2}$:

$$\langle A, B \rangle = \operatorname{trace}(A^\top B) = \sum_{d_1 \in \mathbb{N}_{D_1}} \sum_{d_2 \in \mathbb{N}_{D_2}} a_{d_1 d_2} b_{d_1 d_2} \quad (1)$$

where $\cdot^\top$ indicates matrix transposition[1]. The corresponding Hilbert-Frobenius norm is $\|A\| = \sqrt{\langle A, A\rangle}$.

In this work we are concerned with a multivariate time series $x_t \in \mathbb{R}^D$ following a Vector Autoregressive (VAR) model of order $P$:

$$x_t = \Psi_1 x_{t-1} + \Psi_2 x_{t-2} + \cdots + \Psi_p x_{t-p} + e_t, \qquad (2)$$

where for any $p \in \mathbb{N}_P$, $\Psi_p \in \mathbb{R}^{D\times D}$ and the innovation $e_t \in \mathbb{R}^D$ is a zero-mean serially uncorrelated stationary process.

*Remark 1.* We do not consider here the case where (2) includes a deterministic trend; note, however, that the convex approach that we consider in the following can be easily adapted to deal also with this case.

### *2.1 Nonstationarity and Cointegration*

Recall that $x_t$ is called *unit-root stationary* if all the zeros of the univariate polynomial matrix of degree $P$

$$\mathcal{P}(z) = I - \Psi_1 z - \Psi_2 z^2 - \cdots - \Psi_p z^p$$

are outside the unit circle, see e.g. Tsay [2005]. If $\det(\mathcal{P}(1)) = 0$, then $x_t$ is *unit-root nonstationary*. Following Johansen [1992] we call $x_t$ *integrated of order* $R$ (or simply $\mathcal{I}(R)$) if the process $\Delta^R x_t = x_t - x_{t-R}$ is stationary whereas for any $r \in \mathbb{N}_{R-1}$, the process $\Delta^r x_t$ is not. A stationary process is referred to as $\mathcal{I}(0)$. Finally we say that a $\mathcal{I}(R)$ process $x_t$ is *cointegrated* if there exists at least a *cointegrating vector* $\lambda \in \mathbb{R}^D$ such that the scalar process $\lambda^\top x_t$ is $\mathcal{I}(R^*)$ with $R^* < R$. Cointegrated processes were originally introduced in Granger [1981] and Engle and Granger [1987]. Since then they have become popular mostly in theoretical and applied econometrics.

Cointegration has been advocated to explain long-run or equilibrium relationships; for more discussions on cointegration and cointegration tests, see Box and Tiao [1977], Engle and Granger [1987], Stock and Watson [1988], Johansen [1988]. In the following we focus on the situation where $x_t$ is $\mathcal{I}(1)$. This case is the most commonly studied in the literature.

### *2.2 Error Correction Models*

An *error correction model* (ECM) for the VAR($P$) process (2) is Tsay [2005], Lütkepohl [2005]:

$$\Delta x_t = \Pi x_{t-1} + \Phi_1 \Delta x_{t-1} + \Phi_2 \Delta x_{t-2} + \cdots + \Phi_{P-1} \Delta x_{t-P+1} + e_t, \qquad (3)$$

where $\Phi_p = -\sum_{j=p+1}^{P} \Psi_j$ and $\Pi = -\mathcal{P}(1)$. The VAR($P$) model can be recovered from the ECM via:

$$\begin{cases} \Psi_1 &= I + \Pi + \Phi_1, & (4a)\\ \Psi_p &= \Phi_p - \Phi_{p-1},\ p = 2, \ldots, P-1, & (4b)\\ \Psi_P &= -\Phi_{P-1}\ . & (4c) \end{cases}$$

Model-based forecasting is done as in the stationary case. Assume $e_t$ is independent white noise with covariance matrix $\Sigma_e$. It can be shown that the optimal $H-$step forecast at the origin is

$$y_t(H) = \Psi_1 y_t(H-1) + \cdots + \Psi_p y_t(H-P) \qquad (5)$$

where $y_t(j) = y_{t+j}$ for $j \le 0$. The forecast mean square error (MSE) matrix is $\Sigma_y(H) = \Sigma_e + \sum_{h\in\mathbb{N}_{H-1}} \Psi_h \Sigma_e \Psi_h^\top$; notably for unit-root nonstationary processes some entries will approach infinity as $h \to \infty$ Lütkepohl [2005].

### *2.3 Cointegrating Rank*

Note that, under the assumption that $x_t$ is at most $\mathcal{I}(1)$, $\Delta x_t$ is a $\mathcal{I}(0)$ process. The *cointegrating rank* Engle and Granger [1987] is defined as:

$$D^\star = \operatorname{rank}(\Pi)\ . \qquad (6)$$

One can distinguish the following situations depending on its value:

(1) $D^\star = 0$. In this case $\Pi = 0$ and (3) reduces to a VAR($P-1$) model.

(2) $0 < D^\star < D$. The matrix $\Pi$ can be written as

$$\Pi = AB^\top \qquad (7)$$

where $A, B \in \mathbb{R}^{D\times D^\star}$ are full (column) rank. The $D^\star$ linearly independent columns of $B$ are cointegrating vectors; they form a basis for the *cointegrating subspace*. The vector process

$$f_t = B^\top x_t \qquad (8)$$

is stationary and represent deviation from equilibrium. These equations clarify that (3) expresses the change in $x_t$ in terms of the deviations from the equilibrium at time $t-1$ (the term $\Pi x_{t-1} = A f_{t-1}$, called *error correction term*) and previous changes (the terms $\Phi_p \Delta x_{t-p}$, $1 \le p \le P-1$).

(3) $D^\star = D$. In this case $\det(\mathcal{P}(1)) \neq 0$ ($\Pi$ is full rank); $x_t$ is $\mathcal{I}(0)$ and one can study (2) directly.

An example of cointegrated process is given in Figure 2.

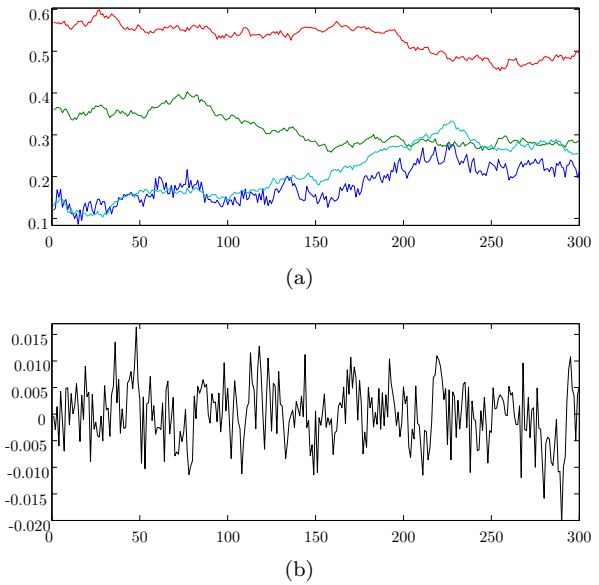


Fig. 1. (a) A realization of the scalar components of a 4-dimensional $\mathcal{I}(1)$ cointegrated process with cointegrating rank 1. (b) The stationary process corresponding to the cointegrating vector $[0.4\ -0.2\ 0.1\ -0.3]$.

[1] In this paper all vectors and matrices are real.

### 2.4 Existing Estimators

Different approaches have been proposed to estimate ECMs based on training data $\{x_t\}_{1\leq t\leq T}$, see [Lütkepohl, 2005, Chapter 7.2]. Define the matrices

$$X_{-p} = [\, x_{P-p+1},\, x_{P-p+2},\, \cdots,\, x_{T-p}\,], \tag{9a}$$
$$\Delta X_{-p} = X_{-p} - X_{-p-1}, \tag{9b}$$
$$\Delta X = \left[\Delta X_{-1}^\top, \Delta X_{-2}^\top, \cdots, \Delta X_{-P+1}^\top\right]^\top, \tag{9c}$$
$$\Phi = [\Phi_1, \Phi_2, \cdots, \Phi_{P-1}], \tag{9d}$$
$$C = \left[X_{-1}^\top, \Delta X^\top\right]^\top, \tag{9e}$$

where $X_{-p} \in \mathbb{R}^{D\times(T-P)}$, $\Delta X_{-p} \in \mathbb{R}^{D\times(T-P)}$, $\Delta X \in \mathbb{R}^{(P-1)D\times(T-P)}$, $\Phi \in \mathbb{R}^{D\times(P-1)D}$ and $C \in \mathbb{R}^{PD\times(T-P)}$. The *Least Squares* (LS) estimator is simply Lütkepohl [2005]:

$$\left[\hat{\Pi}_{LS},\ \hat{\Phi}_{LS}\right] = \Delta X_0 C^\top \left(CC^\top\right)^{-1} . \tag{10}$$

This approach does not keep the decomposition (7) of $\Pi$ into account [2]; in principle, one could perform truncated *singular value decomposition* (SVD) of $\hat{\Pi}_{LS}$ a posteriori to find $B$. However, this practice requires the knowledge of $D^\star$, which is normally not available [3]. When this information is available, an alternative approach consists of Maximum Likelihood Estimation (MLE), which works under the assumption that the innovation is independent white Gaussian noise. Contrary to LS, this approach directly estimates the factors in the representation (7). This leads to a nonconvex multistage algorithm. We refer the reader to [Tsay, 2005, Section 8.6.2] for details.

## 3. ESTIMATION BASED ON CONVEX PROGRAMMING

Recall that for an arbitrary matrix $A \in \mathbb{R}^{D_1\times D_2}$ with rank $R$, the SVD is

$$A = U\Sigma V^\top, \quad \Sigma = \mathrm{diag}(\{\sigma_r\}_{1\leq r\leq R}) \tag{11}$$

where $U \in \mathbb{R}^{D_1\times R}$ and $V \in \mathbb{R}^{D_2\times R}$ are matrices with orthonormal columns, and the singular values $\sigma_r$ satisfy $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_R > 0$. The *nuclear norm* (a.k.a. trace norm or Schatten$-1$ norm) is defined Horn and Johnson [1994] as

$$\|A\|_* = \sum_{r\in\mathbb{N}_R} \sigma_r . \tag{12}$$

The nuclear norm has been used to devise convex relaxations to rank constrained matrix problems Recht et al. [2007], Candès and Recht [2009], Candes et al. [2011]. This parallels the approach followed by estimators like the LASSO (*Least Absolute Shrinkage and Selection Operator*, Tibshirani [1996]) that estimate an high dimensional loading vector $x \in \mathbb{R}^D$ based on the $l_1$-norm

$$\|x\|_1 = \sum_{d\in\mathbb{N}_D} |x_d| . \tag{13}$$

[2] In the literature (10) is sometimes called *unrestricted* LS to emphasize that the model's parameters are not constrained to satisfy any specific structural form.

[3] As we later illustrate in experiments, the spectrum of $\hat{\Pi}_{LS}$ normally does not give a good indication of the actual value of the cointegrating rank.

### 3.1 Main Problem Formulation

Note that, based upon (9), the ECM (3) can be restated in matrix notation as $\Delta X_0 = \Pi X_{-1} + \Phi\Delta X + E$. Our approach now consists of finding estimates $(\hat{\Pi}(\lambda), \hat{\Phi}(\lambda))$ by solving:

$$\min_{\Pi,\Phi}\ \frac{1}{2c}\|\Delta X_0 - \Pi X_{-1} - \Phi\Delta X\|^2 + \lambda\|\Pi\|_* \tag{14}$$

where $c$ is a fixed normalization constant, such as $c = D(T-P)$, and $\lambda$ is a trade-off parameter. When $\lambda = 0$ (14) reduces to the LS estimates (10). For a strictly positive $\lambda$, (14) fits a model to the data by minimizing a trade-off between an error function and a proxy for the cointegrating rank; a positive $\lambda$ reflects the prior knowledge that the process should be cointegrated with cointegrating rank $D^\star < D$.

It can be shown that (12) is the *convex envelope* on the unit ball of the (non-convex) rank function Fazel [2002]. In the present context, this makes (14) the tightest convex relaxation to the problem:

$$\min_{\Pi,\Phi}\ \frac{1}{2}\|\Delta X_0 - \Pi X_{-1} - \Phi\Delta X\|^2 + \lambda\,\mathrm{rank}(\Pi) . \tag{15}$$

The nonconvexity of the latter implies that practical algorithms can only be guaranteed to deliver local solutions of (15) that are rank deficient for sufficiently large $\lambda$. In contrast, the problem (14) is *convex* since the objective is a sum of convex functions Boyd and Vandenberghe [2004]. This implies that any local solution found by a provably convergent algorithm is globally optimal.

Once a solution of (14) is available, one can recover the parameters of the VAR($P$) model (2) based upon (4). Note that here we focus on the case where the error function is expressed in terms of the Hilbert-Frobenius norm; however, alternative norms might also be meaningful. For later reference, note that, when $P = 1$ and $c = 1$, (14) boils down to:

$$\min_{\Pi}\ \frac{1}{2}\|\Delta X_0 - \Pi X_{-1}\|^2 + \lambda\|\Pi\|_* . \tag{16}$$

### 3.2 $l_2$ Smoothing

The nature of the problem makes it difficult to find a solution of (14) for $\lambda \to 0$. Indeed, in practice $X_{-1}$ and $\Delta X$ are often close to singular so that the problem is ill-conditioned. In order to improve numerical stability a possible approach is to add a *ridge penalty* to the objective of (14). That is, for a small user-defined parameter $\mu > 0$, to add:

$$\frac{\mu}{2}\left(\|\Pi\|^2 + \sum_{p\in\mathbb{N}_{P-1}} \|\Phi_p\|^2\right) ; \tag{17}$$

we call the resulting optimization problem the $l_2$-*smoothed formulation*. The idea has a long tradition both in optimization and statistics. Recently it found application in the *Elastic Net* Zou and Hastie [2005]. The Elastic Net

finds an high dimensional loading vector $x$ based on empirical data. The approach consists of replacing the LASSO penalty based on the $l_1-$norm (13) with the composite penalty $\lambda_1\|x\|^2+\lambda_2\|x\|_1$. This strategy aims at improving the LASSO in the presence of dependent variables (which, in fact, leads to ill-conditioning).

In the present context it is easy to see that the solution of the $l_2$ smoothed formulation can be found solving (14) where the data matrices $(\Delta X_0, X_{-1}, \Delta X)$ have been replaced by *lifted* matrices $(\Delta X_0^\mu, X_{-1}^\mu, \Delta X^\mu)$. Consider for simplicity problem (16). The lifted matrices of interest become:

$$\Delta X_0^\mu = [\Delta X_0, O] \text{ and } X_{-1}^\mu = [X_{-1}, \sqrt{\mu} I] \tag{18}$$

where $O, I \in R^{D\times D}$ and $O$ is a matrix of zeros. For this reason in the following we will uniquely discuss an algorithm for the non-smoothed case; it is understood that a solution for the smoothed formulation can be found simply replacing the data matrices.

### *3.3 Dual Representation and Duality Gap*

In this section, for simplicity of exposition, we restrict ourselves to the primal problem (16) and derive its dual representation. The *Fenchel conjugate* Rockafellar [1974] of $\|X\|_*$ is:

$$f(A) = \max_X \langle A, X\rangle - \|X\|_* \ . \tag{19}$$

Recall that the dual norm of the nuclear norm is the *spectral norm*, denoted as $\|\cdot\|_2$; for a generic matrix $A$ with SVD (11) such norm is defined by

$$\|A\|_2 = \sigma_1 \ . \tag{20}$$

It is a known fact that the conjugate of a norm is the indicator function of the dual norm unit ball Boyd and Vandenberghe [2004]. In the present context this fact reads as follows.

*Lemma 2.*

$$f(A) = \begin{cases} 0, \text{ if } \|A\|_2 \le 1 \\ \infty, \text{ otherwise } . \end{cases} \tag{21}$$

With reference to (16), it can be shown that strong duality holds; the dual problem can be obtained as follows:

$$\begin{aligned}
&\min_\Pi \{\tfrac{1}{2}\|\Delta X_0 - \Pi X_{-1}\|^2 + \lambda\|\Pi\|_*\} = \\
&\min_\Pi \max_\Lambda \{\langle \Lambda, \Delta X_0 - \Pi X_{-1}\rangle - \tfrac{1}{2}\langle\Lambda,\Lambda\rangle + \lambda\|\Pi\|_*\} = \\
&\max_\Lambda \min_\Pi \{\langle \Lambda, \Delta X_0 - \Pi X_{-1}\rangle - \tfrac{1}{2}\langle\Lambda,\Lambda\rangle + \lambda\|\Pi\|_*\} = \\
&\max_\Lambda \min_\Pi \{\langle \Lambda, \Delta X_0\rangle - \tfrac{1}{2}\langle\Lambda,\Lambda\rangle - \langle \Lambda X_{-1}^\top, \Pi\rangle + \lambda\|\Pi\|_*\} = \\
&\overset{\text{by Lemma 2}}{=} \max_\Lambda \{\langle \Lambda, \Delta X_0\rangle - \tfrac{1}{2}\langle\Lambda,\Lambda\rangle \ : \ \|\Lambda X_{-1}^\top\|_2 \le \lambda\} \ .
\end{aligned} \tag{22}$$

Additionally, as it results clear from the second line of (22), the primal and dual solution are related as follows:

$$\hat\Lambda = \Delta X_0 - \hat\Pi X_{-1} \ . \tag{23}$$

This fact can be readily used to derive an optimality certificate based on the *duality gap*, i.e. the difference of values between the objective functions of the dual and primal problems.

*Remark 3.* Note that the solution $\hat\Lambda$ of the dual problem in the last line of (22) corresponds to the projection of $\Delta X_0$ onto the convex set $\mathcal{S} = \{\Lambda \ : \ \|\Lambda X_{-1}^\top\|_2 \le \lambda\}$.

## 4. ALGORITHM

In order to find a solution $(\hat\Pi(\lambda), \hat\Phi(\lambda))$ corresponding to a fixed value of $\lambda$ one could restate (14) as a semidefinite programming problem (SDP) and rely on general purpose solvers such as SeDuMi Sturm [1999]. Alternatively, it is possible to use a modelling language like CVX Grant and Boyd [2010] or YALMIP Löfberg [2004]. However these approaches are practically feasible only for relatively small problems. In contrast, the iterative scheme that we present next can be easily implemented and scales well with the problem size. Additionally, the approach can be conveniently used through warm-starting to compute solutions corresponding to nearby values of $\lambda$. The procedure, detailed in the algorithmic environment below, can be shown to be a special instance of the *fast iterative shrinkage-thresholding algorithm* (FISTA) proposed in Beck and Teboulle [2009]; therefore it inherits its favorable convergence properties, see Beck and Teboulle [2009]. We call it Cointegrated VAR($P$) via FISTA (Co-VAR($P$)-FISTA).

**Algorithm: CoVAR($P$)-FISTA**

**Input:** $X_{-p};\ \Delta X_{-p},\ p = 0, 1, \dots, P-1;$

**Initialize:**

$$\Pi_0 = \Pi_1^*;\ \Phi_{0\,p} = \Phi_{1\,p}^*,\ p \in \mathbb{N}_{P-1};$$

$$t_1 = 1;\quad L = 1/c\left\|CC^\top\right\|_2 \quad \text{(see (9e))}$$

**Iteration** $k \ge 1$**:**

$$A_k = \Pi_k^* X_{-1} + \sum_{p\in\mathbb{N}_{P-1}} \Phi_{kp}^* \Delta X_{-p} - \Delta X_0 \tag{24a}$$

$$\Phi_{k\,p} = \Phi_{k\,p}^* - \frac{1}{Lc} A_k \Delta X_{-p}^\top,\ p \in \mathbb{N}_{P-1} \tag{24b}$$

$$C_k = \Pi_k^* - \frac{1}{Lc} A_k X_{-1}^\top \tag{24c}$$

$$\Pi_k = D_{\frac{\lambda}{L}}(C_k) \quad \text{(see (25))} \tag{24d}$$

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2},\ r_{k+1} = \left(\frac{t_k - 1}{t_{k+1}}\right) \tag{24e}$$

$$\Pi_{k+1}^* = \Pi_k + r_{k+1}(\Pi_k - \Pi_{k-1}) \tag{24f}$$

$$\Phi_{k+1\,p}^* = \Phi_{k\,p} + r_{k+1}(\Phi_{k\,p} - \Phi_{k-1\,p}^*),\ p \in \mathbb{N}_{P-1} \tag{24g}$$

**Return:** $\Pi_k,\ \Phi_{k\,1},\ \Phi_{k\,2}\ , \cdots, \Phi_{k\,P-1}$

The approach is essentially a *forward-backward splitting technique* (see Bauschke and Combettes [2011] and reference therein) which is accelerated to reach the optimal rate of convergence in the sense of Nesterov [1983, 2003].

The procedure is based on two sets of working variables:

$$(\Pi_k,\ \Phi_{k\,1},\ \Phi_{k\,2}\ , \cdots, \Phi_{k\,P-1})$$

and

$$(\Pi_k^*,\ \Phi_{k\,1}^*,\ \Phi_{k\,2}^*\ , \cdots, \Phi_{k\,P-1}^*) \ .$$

Equations (24) from a to d correspond to a forward step in the first set of variables conditioned on the variables in the second set; the step size is determined by the Lipschitz constant $L$; equation (24d) represent the backward step. It amounts at evaluating at the current estimate the *singular value shrinkage operator* Cai et al. [2010] defined, for of a

matrix $A$ with SVD (11), as:

$$D_\tau(A) = U\Sigma_+ V^\top, \quad \Sigma_+ = \mathrm{diag}(\{\max(\sigma_r, \tau)\}_{1\leq r\leq R}) \ . \tag{25}$$

$D_\tau(\cdot)$ is the *proximity operator* Bauschke and Combettes [2011] of the nuclear norm function. Equation (24e) defines the updating constant $r_k$ based upon the *estimate sequence* $t_k$ Nesterov [1983, 2003]. Finally, equations (24) from g to i update the second set of variables based upon the variables in the fist set.

The approach requires to set an appropriate termination criterion. A sensible idea, which we follow in experiments, is to stop when the duality gap corresponding to the current estimate is smaller that a predefined threshold.

## 5. CONTINUOUS SHRINKAGE AND COINTEGRATING RANK

### 5.1 Regularization Path

By continuously varying the regularization parameter $\lambda$ in (14) one obtains an entire set of solutions, denoted as $\{(\hat{\Pi}(\lambda), \hat{\Phi}(\lambda))\}_\lambda$, and called *regularization path*. In general, continuous shrinkage is known to feature certain favorable properties. In particular, the paths of estimators like the LASSO are known to be more stable than those of inherently discrete procedures such as *Subset Selection* Breiman [1996].

In the present context the path begins at $\lambda = 0$ and terminates at the least value $\lambda_{\max}$ leading to $\hat{\Pi}(\lambda_{\max}) = 0$. For problem (16), in particular, such value is

$$\lambda_{\max} = \|\Delta X_0 X_{-1}^\top\|_2 \tag{26}$$

as one can see in light of (23) and Remark 3.

### 5.2 Estimation of the Cointegrating Rank

Denote by $\{\hat{\sigma}_r(\lambda)\}_{1\leq r\leq R}$ the spectrum of $\hat{\Pi}(\lambda)$. For $a > 1$ consider the vector $\tilde{m} \in \mathbb{R}^D$ defined entry-wise by [4]:

$$\tilde{m}_d = \left( \int_{-\infty}^{\log_a(\lambda_{\max})} \hat{\sigma}_d^2(a^t) a^t dt \right)^{1/2} . \tag{27}$$

We call $m$, the vector obtained from the inverse order statistics [5] of $\tilde{m}$, the *path spectrum*. Note that the integral in (27) weights more the area of the path which corresponds to shrunk singular values. The rationale behind this is simple: those singular values that survive the shrinking are likely to be the most important ones. The path spectrum can be used to define an estimator for the cointegrating rank. In particular, one can take

$$\hat{D}_\tau^\star(m) = \arg\min_{d\in\mathbb{N}^D} \left\{ f(d) = \left( \frac{\sum_{i\in\mathbb{N}_d} m_i^2}{\sum_{i\in\mathbb{N}_D} m_i^2} - \tau \right) : \ f(d) > 0 \right\} \tag{28}$$

where $0 < \tau < 1$ is a predefined threshold. Note that this estimator is independent on $\lambda$. This is a desirable feature: setting an appropriate value for the regularization parameter is known to be a difficult task.

[4] In experiments we always consider $a = 10$.

[5] I.e., $m = [\tilde{m}_{(1)}, \tilde{m}_{(2)}, \cdots, \tilde{m}_{(d)}]$ is obtained from $\tilde{m}$ sorting its entries in decreasing order. Note that, normally, $\tilde{m}$ already coincide with its inverse order statistics $m$.

In practice the regularization path is evaluated at discrete values of $\lambda$. Therefore the integrals are replaced by their Monte Carlo estimates. In computing the path we actually begin with $\lambda_{\max}$ (recall that $\hat{\Pi}(\lambda_{\max}) = 0$) and proceed backwards computing the solutions corresponding to logarithmically spaced values of the parameter. At each step we initialize the algorithm of Section 4 with the previous solution.

## 6. EXPERIMENTS

To test the performance of the proposed estimator we considered cointegrated systems with a triangular representation Phillips [1991]. More specifically we generated realizations of a $D-$dimensional cointegrated process $x_t$ with $M$ cointegrating vectors by the following model [6]:

$$x_{it} = \begin{cases} \sum_{j=iM+1}^{(i+1)M} x_{jt} + e_{it}, \text{ if } i = 1, 2, \ldots, M \\ x_{it-1} + e_{it}, \text{ if } i = M+1, M+2, \ldots, D \ . \end{cases} \tag{29}$$

In all the cases $e_t \in \mathbb{R}^D$ is a Gaussian white noise with mean zero and covariance matrix $(5e-3)I_D$. The process is observed in noise: we actually used as training data $N$ successive time steps from a realization of $\tilde{x}_t = x_t + w_t$ where $w_t$ is zero-mean Gaussian white noise with covariance matrix $\eta^2 I_D$. For all the cases we took $P = 1$ and computed the path corresponding to an $l_2$-smoothed formulation with smoothing parameter $\mu$. We compared $\hat{D}_{0.8}^\star(m)$ (see (28)) against the naive estimator $\hat{D}_{0.8}^\star(\hat{\sigma}_{LS})$ where $\hat{\sigma}_{LS}$ are the singular values of $\hat{\Pi}_{LS}$. Figure 2 refers to an experiment with $D = 72,\ M = 8,\ N = 400,\ \eta = 0.002$ and $\mu = 0.01$. In Table 1 we reported the average estimated cointegrating rank (with standard deviation in parenthesis) over 20 random experiments performed for two different set of values of $D,\ M,\ N,\ \eta$ and $\mu$.

Table 1. Estimated cointegrating ranks in randomized experiments

| $D = 16,\ M = 2,\ N = 60,\ \eta = 0.001,\ \mu = 0.1$ | |
|---|---|
| $\hat{D}_{0.8}^\star(m)$ | $\hat{D}_{0.8}^\star(\hat{\sigma}_{LS})$ |
| 2.3(0.5) | 4.35(0.5) |

| $D = 72,\ M = 8,\ N = 400,\ \eta = 0.002,\ \mu = 0.01$ | |
|---|---|
| $\hat{D}_{0.8}^\star(m)$ | $\hat{D}_{0.8}^\star(\hat{\sigma}_{LS})$ |
| 7.4(1.9) | 14.3(0.8) |

## 7. CONCLUSIONS

We presented a novel approach, based on a convex program, for the analysis of cointegrated VAR processes from observational data. We proposed to compute solutions via a scalable iterative scheme; used in combination with warm starting this algorithm can be conveniently employed to compute the entire regularization path. At each step one can rely on the duality gap as an optimality certificate. The regularization path offers indication for the actual value of the cointegrating rank and can be used to define estimators for the latter. An important advantage is

[6] From (29) it is easy to recover the error correction model representation.

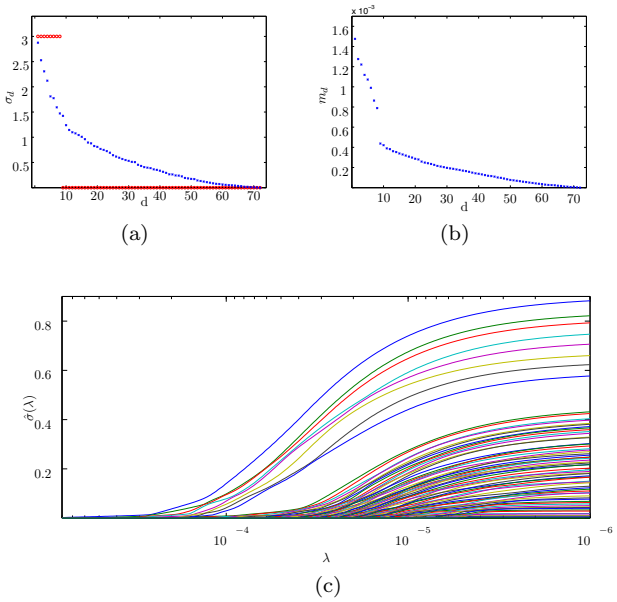


Fig. 2. (a) The singular values of the true $\Pi$ and $\hat{\Pi}_{LS}$; note that the LS solution does not give an indication of the cointegrating rank. (b) The path spectrum; note the gap after $d = 8$. (c) the regularization path.

that the approach does not require to fix a value for the regularization parameter in (14). This is known to be a difficult task, especially when the goal of the analysis is model selection rather than low prediction errors.

## REFERENCES


H.H. Bauschke and P.L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer Verlag, 2011.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

G.E.P. Box and G.C. Tiao. A canonical analysis of multiple time series. *Biometrika*, 64(2):355, 1977.

S.P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

L. Breiman. Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24(6):2350–2383, 1996.

J.F. Cai, E.J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

E.J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of ACM*, 58(3):Article 11, 37p., 2011.

R.F. Engle and C.W.J. Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: Journal of the Econometric Society*, pages 251–276, 1987.

M. Fazel. *Matrix rank minimization with applications*. PhD thesis, Elec. Eng. Dept., Stanford University, 2002.

C.W.J. Granger. Some properties of time series data and their use in econometric model specification. *Journal of econometrics*, 16(1):121–130, 1981.

M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. `http://cvxr.com/cvx`, May 2010.

R.A. Horn and C.R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1994.

S. Johansen. Statistical analysis of cointegration vectors. *Journal of economic dynamics and control*, 12(2-3):231–254, 1988.

S. Johansen. A representation of vector autoregressive processes integrated of order 2. *Econometric theory*, 8 (02):188–202, 1992.

J. Löfberg. Yalmip : A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004. URL `http://users.isy.liu.se/johanl/yalmip`.

H. Lütkepohl. *New introduction to multiple time series analysis*. Springer, 2005.

N. F. Møller and P. Sharp. Malthus in cointegration space: A new look at living standards and population in pre-industrial england. Discussion Papers 08-16, University of Copenhagen. Department of Economics, July 2008. URL `http://ideas.repec.org/p/kud/kuiedp/0816.html`.

Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(\frac{1}{k^2})$. In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983, 1983.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Pub, 2003.

J.E. Payne and B.T. Ewing. Population and economic growth: a cointegration analysis of lesser developed countries. *Applied economics letters*, 4(11):665, 1997.

P.C.B. Phillips. Optimal inference in cointegrated systems. *Econometrica: Journal of the Econometric Society*, pages 283–306, 1991.

B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52:471–501, 2007.

R.T. Rockafellar. *Conjugate duality and optimization*, volume 16. Society for Industrial Mathematics, 1974.

S. Sharma. Persistence and stability in city growth. *Journal of Urban Economics*, 53(2):300–320, 2003.

J.H. Stock and M.W. Watson. Testing for common trends. *Journal of the American statistical Association*, pages 1097–1107, 1988.

J.F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11(1):625–653, 1999.

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

R.S. Tsay. *Analysis of financial time series*, volume 543. Wiley-Interscience, 2005.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. Ser. B*, 67(2): 301–320, 2005.