KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

Sparse cointegration

Wilms I, Croux C.

KBI_1423

Sparse cointegration

Ines Wilms*, Christophe Croux

Faculty of Economics and Business, KU Leuven, Belgium

Abstract. Cointegration analysis is used to estimate the long-run equilibrium relations between several time series. The coefficients of these long-run equilibrium relations are the cointegrating vectors. In this paper, we provide a sparse estimator of the cointegrating vectors. The estimation technique is sparse in the sense that some elements of the cointegrating vectors will be estimated as zero. For this purpose, we combine a penalized estimation procedure for vector autoregressive models with sparse reduced rank regression. The sparse cointegration procedure achieves a higher estimation accuracy than the traditional Johansen cointegration approach in settings where the true cointegrating vectors have a sparse structure, and/or when the sample size is low compared to the number of time series. We also discuss a criterion to determine the cointegration rank and we illustrate its good performance in several simulation settings. In a first empirical application we investigate whether the expectations hypothesis of the term structure of interest rates, implying sparse cointegrating vectors, holds in practice. In a second empirical application we show that forecast performance in high-dimensional systems can be improved by sparsely estimating the cointegration relations.

Keyswords. Adaptive lasso; Penalized estimation; Reduced rank regression; Sparse estimation; Vector error correcting model

*Corresponding author: Faculty of Economics and Business, KU Leuven, Naamsestraat 69, B-3000 Leuven, Belgium. E-mail address: ines.wilms@kuleuven.be.

1 Introduction

High-dimensional datasets containing thousands of economic time series are commonly available and accessible at reasonable cost (Stock and Watson, 2002; Clements and Galvao, 2008; Fan et al., 2011). The aim of this paper is to develop a cointegration technique for high-dimensional time series. In a cointegration analysis, long-run equilibrium relations, often implied by economic theory, are estimated. In financial economics, for instance, cointegration analysis is used to investigate whether the expectations hypothesis of the term structure of interest rates (EHT) holds in practice. The Vector Error Correcting Model (VECM) (see e.g. Lutkepohl, 2007) is used to estimate and test for the cointegration relationships. Various approaches to test for cointegration are existing (see among others Engle and Granger, 1987; Phillips and Ouliaris, 1990), among which the system cointegration test of Johansen (1988) has become most popular.

The conventional Johansen system cointegration approach has, however, some limitations. In high-dimensional settings, where the number of time series is large compared to the sample size, the estimation imprecision will be large. Johansen's approach is based on the estimation of a Vector AutoRegressive (VAR) model and a canonical correlation analysis. A drawback of the VAR model is that the number of parameters increases quadratically with the number of included time series. Consequently, regression parameters are estimated inaccurately if only a limited number of observations is available. When the number of time series exceeds the sample size, Johansen's approach can not even be applied.

In this paper, we introduce a penalized maximum likelihood approach to estimate the cointegrating vectors in a sparse way, i.e. some of its components are estimated as exactly zero. Sparse estimation techniques show good performance in various fields, such as, for instance, economics (e.g. Fan et al., 2011), econometrics (e.g. Caner and Zhang, 2014), or macro-economics (e.g. Korobilis, 2013). A sparse cointegration approach is useful for several reasons. First, a sparse approach is justified if economic theory implies sparsity in the cointegrating vectors (as is the case for the EHT, see e.g. Engsted and Tanggaard, 1994). Secondly, a sparse approach facilitates model interpretation since only a limited number

of variables, those corresponding to the non-zero coefficients, enter the estimated long-run equilibrium relations. Thirdly, sparsity improves forecast performance through variance reduction. Lastly, the sparse cointegration technique, in contrast to Johansen's method, can be applied when the number of time series exceeds the sample size. We show in a simulation study that the sparse cointegration technique significantly outperforms Johansen's approach when the cointegrating vectors have a sparse structure or when the number of time series is large compared to the sample size.

We apply the sparse cointegration technique on a financial and macro-economic dataset. In the first empirical application, we investigate whether the expectations hypothesis of the term structure of interest rates (EHT) holds in practice. Previous research on the validity of the EHT reports evidence in support of the theory at the short end of the term structure (e.g. Hall et al., 1992; Lasak and Velasco, 2014). The theory is generally rejected at the longer end (e.g. Shea, 1992; Zhang, 1993; Carstensen, 2003). We test the cointegration implications linked to the EHT for five US interest rates. Using the sparse cointegration technique, we find evidence in favor of the zero-sum restriction (i.e. for each cointegrating vector, the sum of the cointegration coefficients should be equal to zero). In a second empirical application, we use the VECM to perform a forecast exercise in a high-dimensional setting containing a large number of industrial production time series. We show that sparsely estimating the cointegrating vectors leads to an improvement in forecast performance.

Cointegration analysis in high-dimensions has received little attention in previous research. Large Vector Autoregressive Models, containing a high number of time series relative to the sample size, have been considered extensively. Common approaches are, among others, Dynamic Factor Models (e.g. Stock and Watson, 2002), Bayesian VAR Models (e.g. Banbura et al., 2010) or Reduced-Rank VAR Models (e.g. Carriero et al., 2011; Bernardini and Cubadda, 2014). Typically, these authors do not account for cointegration. Instead, the time series are either transformed in order to achieve stationarity (e.g. Bernardini and Cubadda, 2014) or the (non)-stationarity of the time series is accounted for in the prior distribution of the autoregressive parameters (e.g. Banbura et al., 2010). Few studies, e.g.

Strachan (2003) or Koop et al. (2006), do account for cointegration by using a Bayesian method for obtaining estimates of the cointegrating vectors. These Bayesian approaches, in contrast to the sparse cointegration approach discussed in this paper, do not perform variable selection and require prior specification.

The remainder of this article is structured as follows. We describe the sparse cointegration method in Section 2. Section 3 provides more details on the algorithm. Section 4 discusses the Rank Selection Criterion (Bunea et al., 2011) to determine the cointegration rank. Section 5 presents the results of a simulation study, Section 6 discusses the findings on the empirical applications. Finally, Section 7 concludes.

2 Penalized Maximum Likelihood

Let $\mathbf{y}_t$ be a q-dimensional multivariate time series, where $\mathbf{y}_t$ is $I(1)$. We assume that $\mathbf{y}_t$ follows a VAR(p) model. Any p^{th} order VAR model can be re-written in vector error correcting (VECM) representation (Hamilton, 1991) as follows

$$\boldsymbol{\Delta}\mathbf{y}_t = \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i \boldsymbol{\Delta}\mathbf{y}_{t-i} + \boldsymbol{\Pi}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \qquad t = 1, \ldots, T \tag{1}$$

where $\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_{p-1}$ are $q \times q$ matrices containing short-run effects, $\boldsymbol{\Pi}$ is a $q \times q$ matrix of rank r, $0 \leq r \leq q$ and $\boldsymbol{\varepsilon}_t$ is assumed to follow a $N_q(\mathbf{0}, \boldsymbol{\Sigma})$.

If we can express $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ $q \times r$ matrices of full column rank r, with $0 < r < q$, then the linear combinations given by $\boldsymbol{\beta}'\mathbf{y}_t$ are stationary and $\mathbf{y}_t$ is said to be cointegrated with cointegration rank r. The cointegrating vectors are the columns of $\boldsymbol{\beta}$ and the adjustment coefficients the elements of $\boldsymbol{\alpha}$.

We estimate the model parameters in a penalized maximum likelihood framework. It is convenient to rewrite model (1) in matrix notation:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Gamma} + \mathbf{Z}\boldsymbol{\Pi}' + \mathbf{E} \tag{2}$$

where $\mathbf{Y} = (\boldsymbol{\Delta}\mathbf{y}_{p+1}, \ldots, \boldsymbol{\Delta}\mathbf{y}_T)'$; $\mathbf{X} = (\boldsymbol{\Delta}\mathbf{X}_{p+1}, \ldots, \boldsymbol{\Delta}\mathbf{X}_T)'$ with $\mathbf{X}_t = (\boldsymbol{\Delta}\mathbf{y}'_{t-1}, \ldots, \boldsymbol{\Delta}\mathbf{y}'_{t-p+1})'$; $\mathbf{Z} = (\mathbf{y}_p, \ldots, \mathbf{y}_{T-1})'$; $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_{p-1})'$; and $\mathbf{E} = (\boldsymbol{\varepsilon}_{p+1}, \ldots, \boldsymbol{\varepsilon}_T)'$. Consider the penalized

negative log-likelihood, given by

$$\mathcal{L}(\mathbf{\Gamma}, \mathbf{\Pi}, \mathbf{\Omega}) = \frac{1}{T}\text{tr}\Big((\mathbf{Y}-\mathbf{X}\mathbf{\Gamma}-\mathbf{Z}\mathbf{\Pi}')\mathbf{\Omega}(\mathbf{Y}-\mathbf{X}\mathbf{\Gamma}-\mathbf{Z}\mathbf{\Pi}')'\Big) - \log|\mathbf{\Omega}| + \lambda_1 P_1(\boldsymbol{\beta}) + \lambda_2 P_2(\mathbf{\Gamma}) + \lambda_3 P_3(\mathbf{\Omega}), \tag{3}$$

with $\text{tr}(\cdot)$ denoting the trace, $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$, and P_1, P_2 and P_3 three penalty functions.

We use L_1 or Lasso penalization (Tibshirani, 1996) on the cointegrating vectors

$$P_1(\boldsymbol{\beta}) = \sum_{i=1}^{q}\sum_{j=1}^{r}|\beta_{ij}|. \tag{4}$$

As an extension, we also consider the Adaptive Lasso (Zou, 2006)

$$P_1(\boldsymbol{\beta}) = \sum_{i=1}^{q}\sum_{j=1}^{r} w_{ij}|\beta_{ij}|, \tag{5}$$

with weights w_{ij}. The weights $\hat{w}_{ij}$ are computed as the inverse of the Lasso solution $\hat{w}_{ij} = 1/\hat{\beta}_{ij}^{lasso}$, for $\hat{\beta}_{ij}^{lasso} \neq 0$. The Adaptive Lasso enjoys the oracle property (consistent for variable selection), whereas the Lasso does not (Zou, 2006).

For the short-run effects $\mathbf{\Gamma}$, we use L_2 or Ridge penalization (Hoerl and Kennard, 1970)

$$P_2(\mathbf{\Gamma}) = \sum_{i=1}^{q}\sum_{j=1}^{q}\sum_{k=1}^{p-1}\gamma_{ijk}^2, \tag{6}$$

with γ_{ijk} the $(i,j)th$ element of $\mathbf{\Gamma}_k$. The L_1 penalty shrinks parameter estimates towards zero and sets some to exactly zero. Contrary to the L_1 penalty, the L_2 penalty only shrinks parameter estimates towards zero. We use an L_2 penalty for the short-run effects since this is computationally less expensive and we only require sparsity in the cointegrating vectors. Note that using the ridge penalty, estimation remains feasible if the number of time series exceeds the sample size.

Finally, we use L_1 penalization for the off-diagonal elements of the inverse of the error covariance matrix, $\mathbf{\Omega}$,

$$P_3(\mathbf{\Omega}) = \sum_{k\neq k\prime}|\Omega_{kk'}|. \tag{7}$$

The aim is to select $\mathbf{\Gamma}, \mathbf{\Pi}, \mathbf{\Omega}$ so as to minimize (3) subject to the constraint

$$\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}',$$

with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ $q \times r$ matrices of full column rank r. The matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are not uniquely defined. Section 3 provides more details on the normalization conditions we impose. For the unpenalized case (i.e. $\lambda_1 = 0$, $\lambda_2 = 0$ and $\lambda_3 = 0$), the objective function (3) boils down to the one introduced by Johansen (1988). The unpenalized case can be solved either by using an iterative algorithm or by using the closed-form expressions documented in Johansen (1988).

3 Algorithm

To find the minimum of the penalized negative log-likelihood in (3), we iteratively solve for $\boldsymbol{\Gamma}$ conditional on $\boldsymbol{\Pi}, \boldsymbol{\Omega}$; for $\boldsymbol{\Pi}$ conditional on $\boldsymbol{\Gamma}, \boldsymbol{\Omega}$; and for $\boldsymbol{\Omega}$ conditional on $\boldsymbol{\Gamma}, \boldsymbol{\Pi}$.

Solving for $\boldsymbol{\Gamma}$ conditional on $\boldsymbol{\Pi}, \boldsymbol{\Omega}$. When $\boldsymbol{\Pi}$ and $\boldsymbol{\Omega}$ are fixed, the minimization problem in (3) is equivalent to minimizing

$$\widehat{\boldsymbol{\Gamma}}|\boldsymbol{\Pi}, \boldsymbol{\Omega} = \underset{\boldsymbol{\Gamma}}{\operatorname{argmin}} \quad \frac{1}{T}\operatorname{tr}\Big((\mathbf{Y} - \mathbf{Z}\boldsymbol{\Pi}' - \mathbf{X}\boldsymbol{\Gamma})\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\Pi}' - \mathbf{X}\boldsymbol{\Gamma})'\Big) + \lambda_2 P_2(\boldsymbol{\Gamma}). \tag{8}$$

The above minimization problem is a penalized multivariate regression (see e.g. Rothman et al., 2010) of $(\mathbf{Y} - \mathbf{Z}\boldsymbol{\Pi}')$ on $\mathbf{X}$. We solve this penalized multivariate regression using the ridge penalty, as given in equation (6). The closed-form expression for the estimated short-run dynamics $\widehat{\boldsymbol{\Gamma}}$ is given by

$$\widehat{\boldsymbol{\Gamma}} = \left(\mathbf{X}^{\text{ridge}\prime}\mathbf{X}^{\text{ridge}} + \lambda_2 \mathbf{I}\right)^{-1} \mathbf{X}^{\text{ridge}\prime} y^{\text{ridge}},$$

with

$$y^{\text{ridge}} = \left(\boldsymbol{\Omega}^{1/2} \otimes \mathbf{I_n}\right) \operatorname{vec}\left(\mathbf{Y} - \mathbf{Z}\boldsymbol{\Pi}'\right),$$

where the latter is a vector of length nq containing the stacked values of the time series given in the columns of the matrix $(\mathbf{Y} - \mathbf{Z}\boldsymbol{\Pi}')$, and

$$\mathbf{X}^{\text{ridge}} = (\boldsymbol{\Omega}^{1/2} \otimes \mathbf{I_n})(\mathbf{I_q} \otimes \mathbf{Z}),$$

where $\otimes$ stands for the kronecker product.

Solving for $\boldsymbol{\Pi}$ conditional on $\boldsymbol{\Gamma}, \boldsymbol{\Omega}$. When $\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}$ are fixed, the minimization problem in (3) is equivalent to

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})|\boldsymbol{\Gamma}, \boldsymbol{\Omega} = \underset{\boldsymbol{\alpha}, \boldsymbol{\beta}}{\operatorname{argmin}} \quad \frac{1}{T}\operatorname{tr}\Big((\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma} - \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\alpha}')\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma} - \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\alpha}')'\Big) + \lambda_1 P_1(\boldsymbol{\beta}). \tag{9}$$

The above minimization problem boils down to a penalized reduced rank regression (e.g. Chen and Huang, 2012). For identifiability purposes, we impose the normalization conditions $\boldsymbol{\alpha}'\boldsymbol{\Omega}\boldsymbol{\alpha} = \mathbf{I_r}$. We first estimate $\boldsymbol{\alpha}$ conditional on $\boldsymbol{\beta}$, next we estimate $\boldsymbol{\beta}$ conditional on $\boldsymbol{\alpha}$.

For fixed $\boldsymbol{\beta}$, the minimization problem in (9) reduces to

$$\hat{\boldsymbol{\alpha}}|\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\beta} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \quad \frac{1}{T}\operatorname{tr}\Big((\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma} - \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\alpha}')\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma} - \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\alpha}')'\Big) \quad \text{st. } \boldsymbol{\alpha}'\boldsymbol{\Omega}\boldsymbol{\alpha} = \mathbf{I_r}, \tag{10}$$

which is a weighted Procrustes problem (Lissitz et al., 1976). This weighted Procrustes problem for $\boldsymbol{\alpha}$ can be seen as an unweighted Procrustes problem for $\boldsymbol{\alpha}^\star = \boldsymbol{\Omega}^{\mathbf{1/2}}\boldsymbol{\alpha}$. The solution is

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{\Omega}^{-1/2}\boldsymbol{V}\boldsymbol{U}',$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are obtained from the singular value decomposition of

$$\boldsymbol{\beta}'\mathbf{Z}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma})\boldsymbol{\Omega}^{\mathbf{1/2}} = \mathbf{U}\mathbf{D}\mathbf{V}'.$$

Note that Chen and Huang (2012) only consider the case where $\boldsymbol{\Omega} = \mathbf{I}$, and use a Procrustes problem to solve for $\boldsymbol{\alpha}$. A weighted Procrustes problem takes the covariance structure into account.

For fixed $\boldsymbol{\alpha}$, the minimization problem in (9) reduces to

$$\hat{\boldsymbol{\beta}}|\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\alpha} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \quad \frac{1}{T}\operatorname{tr}\Big((\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma} - \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\alpha}')\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma} - \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\alpha}')'\Big) + \lambda_1 P_1(\boldsymbol{\beta}). \tag{11}$$

Since $\boldsymbol{\alpha}^{\star\prime}\boldsymbol{\alpha}^\star = \mathbf{I_r}$, there exists a matrix $\boldsymbol{\alpha}^{\star\perp}$ with orthonormal columns such that $(\boldsymbol{\alpha}^\star, \boldsymbol{\alpha}^{\star\perp})$ is an orthogonal matrix. Then, with $\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma}$,

$$\begin{aligned}
\operatorname{tr}\Big((\tilde{\mathbf{Y}} - \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\alpha}')\boldsymbol{\Omega}(\tilde{\mathbf{Y}} - \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\alpha}')'\Big) &= ||(\tilde{\mathbf{Y}} - \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\alpha}')\boldsymbol{\Omega}^{\mathbf{1/2}}||^2 \\
&= ||(\tilde{\mathbf{Y}}\boldsymbol{\Omega}^{\mathbf{1/2}} - \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\alpha}^{\star\prime})||^2 \\
&= ||(\tilde{\mathbf{Y}}\boldsymbol{\Omega}^{\mathbf{1/2}} - \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\alpha}^{\star\prime})(\boldsymbol{\alpha}^\star, \boldsymbol{\alpha}^{\star\perp})||^2 \\
&= ||(\tilde{\mathbf{Y}}\boldsymbol{\Omega}^{\mathbf{1/2}}\boldsymbol{\alpha}^\star - \mathbf{Z}\boldsymbol{\beta}||^2 + ||(\tilde{\mathbf{Y}}\boldsymbol{\Omega}^{\mathbf{1/2}}\boldsymbol{\alpha}^{\star\perp})||^2.
\end{aligned}$$

Since the second term on the left-hand-side does not involve $\boldsymbol{\beta}$, the minimization problem reduces to

$$\hat{\boldsymbol{\beta}}|\boldsymbol{\Gamma},\boldsymbol{\Omega},\boldsymbol{\alpha} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \quad \frac{1}{T}\operatorname{tr}\Big((\tilde{\mathbf{Y}}\boldsymbol{\Omega}^{\mathbf{1/2}}\boldsymbol{\alpha}^{\star} - \mathbf{Z}\boldsymbol{\beta})(\tilde{\mathbf{Y}}\boldsymbol{\Omega}^{\mathbf{1/2}}\boldsymbol{\alpha}^{\star} - \mathbf{Z}\boldsymbol{\beta})'\Big) + \lambda_1 P_1(\boldsymbol{\beta}). \tag{12}$$

The above minimization problem is a penalized multivariate regression of $\left(\tilde{\mathbf{Y}}\boldsymbol{\Omega}^{\mathbf{1/2}}\boldsymbol{\alpha}^{\star}\right)$ on $\mathbf{Z}$. We consider both a Lasso penalty, as in equation (4), and an Adaptive Lasso penalty, as in equation (5).

Solving for $\boldsymbol{\Omega}$ conditional on $\boldsymbol{\Gamma}, \boldsymbol{\Pi}$. When $\boldsymbol{\Gamma}$ and $\boldsymbol{\Pi}$ are fixed, the minimization problem in (3) is equivalent to minimizing

$$\widehat{\boldsymbol{\Omega}}|\boldsymbol{\Gamma},\boldsymbol{\Pi} = \underset{\boldsymbol{\Omega}}{\operatorname{argmin}} \quad \frac{1}{T}\operatorname{tr}\Big((\mathbf{Y} - \mathbf{Z}\boldsymbol{\Pi}' - \mathbf{X}\boldsymbol{\Gamma})\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\Pi}' - \mathbf{X}\boldsymbol{\Gamma})'\Big) - \log|\boldsymbol{\Omega}| + \lambda_3 P_3(\boldsymbol{\Omega}). \tag{13}$$

The above minimization problem corresponds to penalized covariance estimation. With the penalty term as given in equation (7), this problem can be solved using the glasso algorithm of Friedman et al. (2008).

We iterate solving minimization problem (8), (9) and (13) until the angle between the estimated cointegration space in two successive iterations is smaller than some tolerance value ϵ (e.g. $\epsilon = 10^{-3}$).

Selection of tuning parameters. We select the tuning parameters λ_1, controlling the penalization on the cointegrating vectors, and λ_2, controlling the penalization of the short-run effects, according to a time series cross-validation approach (Hyndman, 2014), see Appendix B. Since the sparseness structure of each cointegrating vector can be different, we allow the selected sparsity parameter λ_1 to be different for each cointegrating vector. The tuning parameter λ_3, controlling the penalization on the off-diagonal elements of $\boldsymbol{\Omega}$, is selected according to the Bayesian Information Criterion (Friedman et al., 2008).

Starting values. A starting value for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ is required. We start with $\boldsymbol{\Omega} = \mathbf{I_q}$. Starting values for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are obtained by first applying the iterative algorithm with an L_2 penalty

on the cointegrating vectors, initialized by taking every component of $\boldsymbol{\beta}$ equal to one and $\boldsymbol{\Gamma}_k$ (for $k = 1, \ldots, p-1$) the identity matrices.

Unpenalized objective function. The unpenalized case (i.e. $\lambda_1 = 0$, $\lambda_2 = 0$ and $\lambda_3 = 0$) can also be solved using the iterative algorithm. We numerically verified that this iterative procedure and Johansen's closed-form solution yield almost identical results, justifying the use of our iterative procedure to solve objective function (1).

4 Determination of Cointegration Rank

At small finite samples, the asymptotic distribution of Johansen's trace statistic, used to determine the cointegration rank, might poorly approximate the true distribution, resulting in substantial size and power distortions (e.g. Johansen, 2002; Nielsen, 2004; Juselius, 2006; Breitung and Cubadda, 2011). We use an iterative procedure based on the Rank Selection Criterion (RSC) of Bunea et al. (2011) to determine the cointegration rank r. We start with an initial value of the cointegration rank $r_{\text{start}} = q$.

For this initial value, we first obtain $\widehat{\boldsymbol{\Gamma}}$, using the algorithm discussed in Section 3. In a second step, we update our estimate of the cointegration rank. Following Bunea et al. (2011), $\hat{r}$ is given by the number of eigenvalues of the matrix $\widetilde{\mathbf{Y}}'\mathbf{P}\widetilde{\mathbf{Y}}$ that exceed a certain threshold μ:

$$\hat{r} = \max\{r : \lambda_r(\widetilde{\mathbf{Y}}'\mathbf{P}\widetilde{\mathbf{Y}}) \geq \mu\},$$

with $\widetilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Gamma}}$ and $\mathbf{P} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-}\mathbf{Z}'$ the projection matrix onto the column space of $\mathbf{Z}$. Following the recommendation of Bunea et al. (2011), the threshold is set equal to $\mu = 2S^2(q + l)$, with $l = \text{rank}(\mathbf{Z})$ and

$$S^2 = \frac{||\widetilde{\mathbf{Y}} - \mathbf{P}\widetilde{\mathbf{Y}}||_F^2}{Tq - lq},$$

where $||\cdot||_F$ denotes the Frobenius norm for a matrix. We repeat the above procedure using the new value of $\hat{r}$, this until the estimate of the cointegration rank does not change in two successive iterations.

The Rank Selection Criterion provides a consistent estimate of the effective rank of the coefficient matrix $\mathbf{\Pi}$ in the penalized reduced rank regression (Bunea et al., 2011). The consistency results are valid when either the length of the time series or the number of time series grows to infinity. This procedure to determine the rank has almost no computational cost and can also be used when the number of time series is larger than the sample size.

5 Simulation Studies

We conduct a simulation study to evaluate the performance of the penalized ML estimator. The considered data generating process (revised from Cavaliere et al., 2012) is the following VECM:

$$\mathbf{\Delta y}_t = \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{y}_{t-1} + \mathbf{\Gamma}_1\mathbf{\Delta y}_{t-1} + \mathbf{e}_t, \qquad (t = 1, \ldots, T),$$

where the error terms $\mathbf{e}_t$ follow a $N_q(\mathbf{0}, \mathbf{I_q})$ distribution. We set $\mathbf{y}_0 = \mathbf{\Delta y}_0 = \mathbf{0}$. We compare the precision accuracy of the penalized ML algorithm to the maximum likelihood procedure of Johansen (1988).

5.1 Simulation designs. Different simulation designs are considered: (i) low-dimensional designs ($T = 500, q = 4$), and (ii) high-dimensional designs with moderate sample size ($T = 50, q = 11$)[1]. For each design, we consider both sparse and non-sparse simulation settings. Full details on each simulation design can be found in Appendix A. The number of simulations for each setting is $M = 500$.

Low-dimensional designs. The true cointegrating vectors are sparse in the first two simulation settings. The cointegration rank equals $r = 1$, $r = 2$ respectively. In the third simulation setting, the true cointegrating vector is non-sparse and $r = 1$.

High-dimensional designs. In the first two simulation settings, the true cointegrating vectors are sparse. The cointegration rank equals $r = 1$, $r = 4$ respectively. In the third

[1] $q = 11$ time series is the largest number for which the critical values of Johansen's trace statistic are tabulated in Johansen (Chapter 15; 1996) or Osterwald-Lenum (1992). Note that Doornik (1998) provide a response surface approximation to the critical values tabulated by Johansen for q up to at least 15.

simulation setting, the true cointegrating vector is non-sparse and $r = 1$.

5.2 Performance measures. To evaluate the estimation accuracy, we compute for each simulation run m, with $m = 1, \ldots, M$, the angle $\theta^{(m)}(\hat{\boldsymbol{\beta}}^{(m)}, \boldsymbol{\beta})$ between the estimated cointegration space and the true cointegration space. The average angle is then given by

$$\theta(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \frac{1}{M} \sum_{m=1}^{M} \theta^{(m)}(\hat{\boldsymbol{\beta}}^{(m)}, \boldsymbol{\beta}). \tag{14}$$

Furthermore, we evaluate the performance of the Rank Selection Criterion to the trace statistic of Johansen (1988), the Bartlett-corrected trace statistic (Johansen, 2002) and the bootstrap procedure of Cavaliere et al. (2012) in correctly selecting the true cointegration rank.[2] The Bartlett-corrected trace statistic (Johansen, 2002) and bootstrap procedure are used to improve the small sample performance of Johansen's trace statistic. For each method, we record the relative frequencies, over all simulation runs, of the selected ranks.

5.3 Results for the low-dimensional designs. The simulation results on the accuracy of the estimated cointegration space are reported in Table 1. For different values of the adjustment coefficients, we report the average angle (averaged across simulation runs) between the estimated and the true cointegration space. We use a two-sided paired t-test to test equality of the average angle of the sparse estimation method and of Johansen's method.

In the sparse settings, the sparse methods are the best performing. They provide significantly more precise estimates than the Johansen procedure. For almost all values of the adjustment coefficients, the estimation accuracy of the sparse methods is even twice as good as that of Johansen's method. The Sparse Adaptive Lasso provides a more precise estimate of the cointegration space than the Sparse Lasso. In the non-sparse setting, Johansen's method is best performing for low values of the adjustment coefficients. For higher values of a, however, all methods show similar performance. The usage of the sparse procedures does not lead to an lower estimation precision here.

Table 2 reports the results on the determination of the cointegration rank. For reasons of brevity, we only report the results for $a = -0.4$ and $a = -0.8$. In the first sparse design, the

[2] All tests are conducted at the 5% significance level.

Table 1: Low-dimensional designs: $T = 500, q = 4$. Average angle between the estimated and true cointegration space. The results are reported for different values of the adjustment coefficient a. Significant differences, at the 5% significance level, between the sparse method and Johansen's method are in bold.

Method \ a	-0.2	-0.4	-0.6	-0.8
Sparse $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}, r = 1$				
Johansen	0.060	0.032	0.020	0.015
Sparse Lasso	**0.034**	**0.018**	**0.012**	**0.009**
Sparse Adaptive Lasso	**0.011**	**0.004**	**0.003**	**0.002**
Sparse $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}, r = 2$				
Johansen	0.013	0.006	0.004	0.003
Sparse Lasso	**0.007**	**0.003**	**0.002**	**0.002**
Sparse Adaptive Lasso	**0.001**	**0.001**	**0.001**	**0.001**
Non-sparse $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}, r = 1$				
Johansen	0.026	0.013	0.009	0.007
Sparse Lasso	**0.073**	0.013	0.009	0.007
Sparse Adaptive Lasso	**0.077**	0.014	0.009	0.007

Table 2: Low-dimensional designs: $T = 500, q = 4$. Frequency of the estimated cointegration rank $\hat{r} = 0, \ldots, q$ using Johansen's trace statistic, the Bartlett-corrected trace statistic, the bootstrap of Cavaliere et al. (2012) and the Rank Selection Criterion (RSC).

True rank	Method \ $\hat{r}$	0	1	2	3	4	0	1	2	3	4
Sparse α and β				**$a = -0.4$**					**$a = -0.8$**		
$r = 1$	Johansen	0.0	95.8	3.8	0.4	0.0	0.0	95.8	4.0	0.2	0.0
	Bartlett	3.6	83.4	11.8	1.2	0.0	5.8	82.6	10.6	1.0	0.0
	Bootstrap	0.0	96.2	3.4	0.4	0.0	0.0	96.0	3.8	0.2	0.0
	RSC	0.0	91.0	9.0	0.0	0.0	0.0	91.6	8.4	0.0	0.0
Sparse α and β				**$a = -0.4$**					**$a = -0.8$**		
$r = 2$	Johansen	0.0	0.0	96.4	3.6	0.0	0.0	0.0	96.0	3.8	0.2
	Bartlett	2.6	7.2	84.8	5.4	0.0	5.0	5.6	83.2	6.2	0.0
	Bootstrap	0.0	0.0	96.0	3.6	0.4	0.0	0.0	95.4	4.2	0.4
	RSC	0.0	0.0	99.6	0.4	0.0	0.0	0.0	99.8	0.2	0.0
Non-sparse α and β				**$a = -0.4$**					**$a = -0.8$**		
$r = 1$	Johansen	0.0	94.6	5.0	0.4	0.0	0.0	95.6	3.6	0.8	0.0
	Bartlett	0.6	88.8	8.8	1.8	0.0	0.6	91.0	7.8	0.6	0.0
	Bootstrap	0.0	95.6	4.2	0.2	0.0	0.0	96.0	3.4	0.6	0.0
	RSC	0.0	90.4	9.6	0.0	0.0	0.0	91.4	8.6	0.0	0.0

Rank Selection Criterion achieves competitive performance with a rank recovery percentage around 91%. Note that Johansen's method is aimed at controlling size, resulting in a rank recovery percentage around 95% when working with a 5% significance level. In the second sparse design, RSC is the best performing method. It correctly selects the cointegration rank in almost all simulation runs. In the non-sparse design, Johansen's procedure performs best. The rank recovery percentage of RSC remains close to that of Johansen's trace statistic.

5.4 Results for the high-dimensional designs. In these designs, we expect that the advantage of using the sparse procedures becomes much larger. The sample size is small

Table 3: High-dimensional designs: $T = 50, q = 11$. Average angle between the estimated and true cointegration space. Results are reported for different values of the adjustment coefficient a. Significant differences, at the 5% significance level, between the sparse method and Johansen's method are in bold.

Method \ a	-0.2	-0.4	-0.6	-0.8
Sparse $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}, r = 1$				
Johansen	1.203	1.025	0.825	0.672
Sparse Lasso	**0.791**	**0.396**	**0.228**	**0.099**
Sparse Adaptive Lasso	**0.816**	**0.392**	**0.209**	**0.090**
Sparse $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}, r = 4$				
Johansen	0.184	0.101	0.064	0.047
Sparse Lasso	**0.154**	**0.076**	**0.047**	**0.034**
Sparse Adaptive Lasso	**0.152**	**0.070**	**0.042**	**0.033**
Non-sparse $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}, r = 1$				
Johansen	1.203	1.005	0.810	0.656
Sparse Lasso	**0.730**	**0.384**	**0.250**	**0.161**
Sparse Adaptive Lasso	**0.758**	**0.403**	**0.266**	**0.174**

compared to the number of time series, such that the estimation imprecision when using Johansen's approach will become large. The simulation results on the estimation accuracy of the estimated cointegration space are reported in Table 3. In all settings, the sparse procedures indeed significantly outperform Johansen's procedure. Also for the non-sparse design the sparse estimation procedures perform best. The differences are outspoken. Since the Lasso and Adaptive Lasso perform regularization, their good performance is retained in non-sparse high-dimensional settings. Furthermore, the Sparse Lasso and the Sparse Adaptive Lasso show similar performance.

Table 4 reports the results on the determination of the cointegration rank. In all simulation designs, the Rank Selection Criterion does much better than its alternatives. In the first

Table 4: High-dimensional designs: $T = 50, q = 11$. Frequency of the estimated cointegration rank $\hat{r} = 0, \ldots, q$.

True rank	Method \ $\hat{r}$	0	1	2	3	4	5	6	7	8	9	10	11
Sparse α and β								$a = -0.4$					
$r = 1$	Johansen	0.0	0.0	0.0	0.0	0.0	0.2	3.8	6.6	22.2	25.4	25.4	16.4
	Bartlett	7.0	59.0	22.4	7.8	3.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0
	Bootstrap	89.6	9.2	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	RSC	3.8	95.2	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
								$a = -0.8$					
$r = 1$	Johansen	0.0	0.0	0.0	0.0	0.0	0.2	3.6	9.2	22.6	24.0	23.4	17.0
	Bartlett	6.6	52.4	23.2	12.4	4.0	1.4	0.0	0.0	0.0	0.0	0.0	0.0
	Bootstrap	83.2	15.0	1.6	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	RSC	0.0	94.6	5.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Sparse α and β								$a = -0.4$					
$r = 4$	Johansen	0.0	0.0	0.0	0.0	0.4	2.8	24.8	22.6	27.2	12.6	5.8	3.8
	Bartlett	6.8	43.8	23.2	16.2	6.2	2.4	0.8	0.6	0.0	0.0	0.0	0.0
	Bootstrap	75.2	21.6	3.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	RSC	1.8	15.0	30.6	37.2	15.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
								$a = -0.8$					
$r = 4$	Johansen	0.0	0.0	0.0	0.0	0.4	1.8	22.4	28.6	28.0	11.2	5.2	2.4
	Bartlett	2.4	32.6	24.2	19.2	10.6	7.6	2.8	0.4	0.2	0.0	0.0	0.0
	Bootstrap	21.6	44.4	25.6	6.8	1.4	0.0	0.0	0.2	0.0	0.0	0.0	0.0
	RSC	0.0	0.0	0.0	3.2	96.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Non-sparse α and β								$a = -0.4$					
$r = 1$	Johansen	0.0	0.0	0.0	0.0	0.0	0.2	2.6	8.4	21.6	26.8	25.0	15.4
	Bartlett	5.8	56.6	26.2	8.6	2.2	0.6	0.0	0.0	0.0	0.0	0.0	0.0
	Bootstrap	88.6	9.6	1.6	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	RSC	6.4	92.6	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
								$a = -0.8$					
$r = 1$	Johansen	0.0	0.0	0.0	0.0	0.0	0.0	2.4	8.6	21.0	25.6	24.0	18.4
	Bartlett	6.8	50.8	26.2	12.4	2.6	0.8	0.4	0.0	0.0	0.0	0.0	0.0
	Bootstrap	80.4	17.4	2.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	RSC	0.0	96.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

design with $a = -0.8$, for instance, RSC estimates the cointegration rank correctly in 94.6% of the simulation runs, the Bartlett-corrected trace statistic in 52.4%, the bootstrap only in 15.0% and Johansen's trace statistic in 0% of the simulation runs. Due to the severe size distortions in this small sample size design, the rank recovery percentage of Johansen's trace statistic does not improve when working with a significance level of, for instance, 1%. The Bartlett-corrected trace statistic shows a considerable improvement over Johansen's trace statistic. Nevertheless, RSC still outperforms the Bartlett-corrected trace statistic.

When the true cointegration rank becomes higher ($r = 4$ in the second design), the performance of the Rank Selection Criterion becomes sensitive to the strength of the cointegration signal: its rank recovery percentage increases from 15.4% for $a = -0.4$ to 96.8% for $a = -0.8$. However, even then, RSC is still the best performing method.

6 Application

We consider two empirical applications. In the first application on interest rates, economic theory implies sparsity in the cointegrating vectors. Therefore, it is appealing to use the sparse cointegration technique even though standard results from Johansen's system cointegration test are also available. Secondly, we perform a forecasting exercise in a large VECM of industrial production time series.

6.1 The term structure of interest rates. We use the sparse cointegration approach to investigate whether the expectations hypothesis of the term structure of interest rates (EHT) holds in practice. The EHT implies that the long-term interest rate can be expressed as an average of current and market-expected future short-term interest rates plus a constant risk premium:

$$r_t^\tau = \frac{1}{\tau} \sum_{i=0}^{\tau-1} \mathbb{E}_t r_{t+i}^1 + C, \tag{15}$$

where r_t^τ and r_t^1 are the τ-period and one-period interest rates, C is a constant term premium and $\mathbb{E}_t$ is the expectations operator conditional on public information at time t (e.g. Lanne, 2000). We consider q interest rates $r_t^1, r_t^{\tau_2}, \ldots, r_t^{\tau_q}$ with increasing time to maturity

$1, \tau_2, \ldots, \tau_q$. Then equation (15) holds for all pairs of interest rates $\{r_t^1, r_t^{\tau_2}\}$, $\{r_t^1, r_t^{\tau_3}\}$, $\ldots$, $\{r_t^1, r_t^{\tau_q}\}$ and we can write

$$r_t^\tau - r_t^1 = \frac{1}{\tau} \sum_{i=1}^{\tau-1} \sum_{j=1}^{i} \mathbb{E}_t \Delta r_{t+j}^1 + C, \tag{16}$$

with $\Delta r_{t+j}^1 = r_{t+j}^1 - r_{t+j-1}^1$. Since the interest rates are assumed to be $I(1)$, the first differences are stationary and, hence, the right-hand-side of equation (16) is stationary. This implies that the left-hand-side of equation (16) must be stationary as well. There are two cointegration implications linked to the EHT. Firstly, there should be $q-1$ cointegrating vectors in a system with q interest rates of different maturity; or equivalently, one common trend (with the number of common trends $= q - r$). Secondly, the $q-1$ yield spreads between the one-period interest rate and each n-period interest rate span the cointegration space:

$$\begin{bmatrix} 1 & 1 & \ldots & 1 \\ -1 & 0 & \ldots & 0 \\ 0 & -1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & -1 \end{bmatrix}. \tag{17}$$

For each cointegrating vector, the sum of the cointegration coefficients should be equal to zero ("zero-sum restriction"). Rejection of one of both implications would be considered as evidence against the EHT.

We collect monthly data on five US treasury bills with different time to maturity (1, 3, 5, 7 and 10 years), ranging from January 1962 until February 2014 (source: Federal Reserve, United States). Time plots on the interest rates in levels, in first differences and the spreads are presented in Figure 1. A stationarity test of all individual interest rates using the Augmented Dickey-Fuller test confirms that the time series are integrated of order 1.

Cointegration Rank. We estimate the cointegration rank using Johansen's trace statistic and the Rank Selection Criterion discussed in Section 4. Table 5 reports the results on the estimated cointegration rank. For the system with two interest rates (2-IR system),

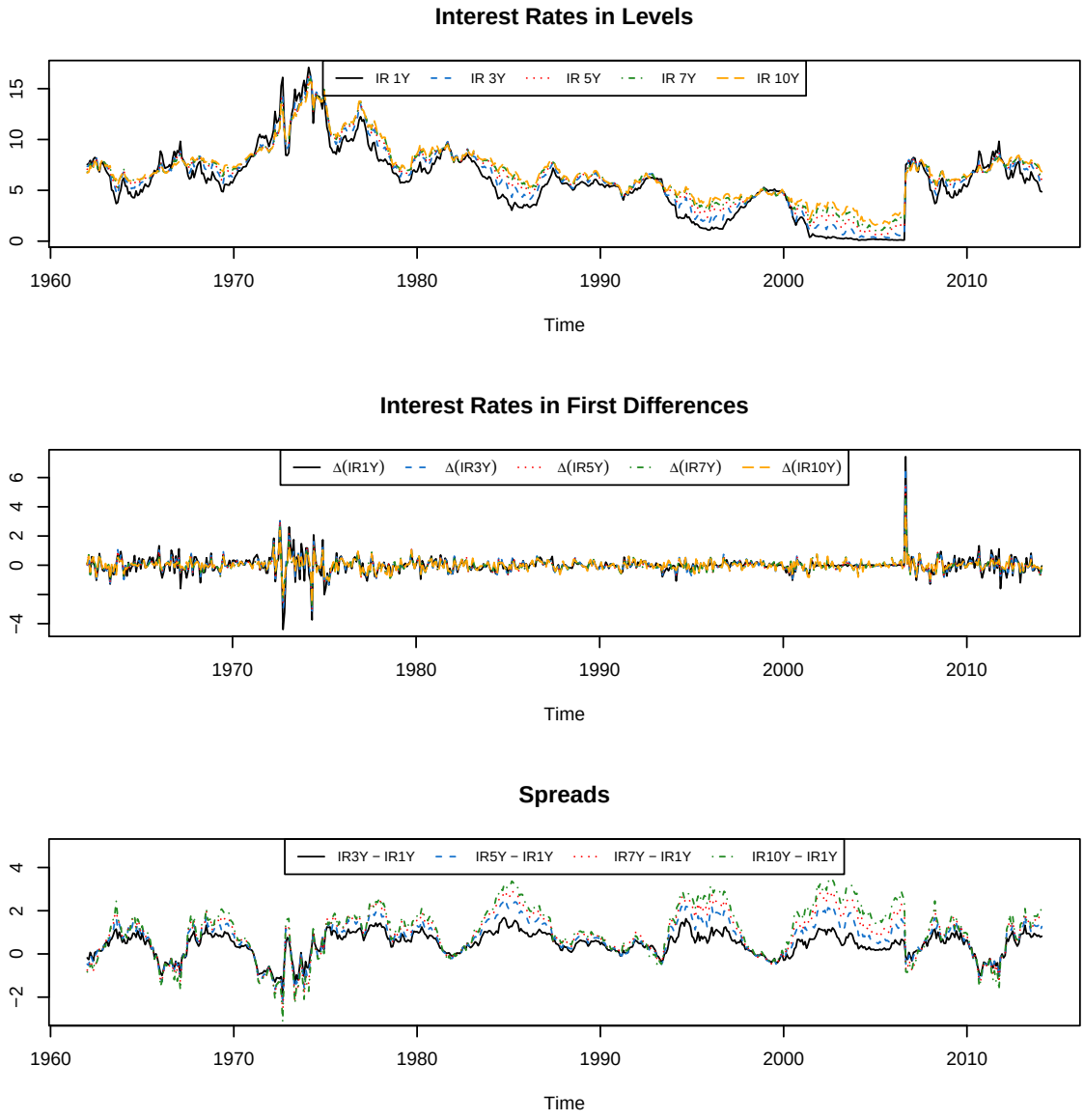

Figure 1: Time plot of the interest rates (US treasury bills, constant maturity: 1-year; 3-year; 5-year; 7-year and 10-year, in % per annum) in levels, in first differences and the spreads. Period January 1962 to February 2014.

Table 5: Estimated cointegration rank using Johansen's trace statistic and the Rank Selection Criterion for the four interest rate systems. The last column reports the cointegration rank implied by the expectations hypothesis.

	Estimated Cointegration Rank[1]		Expectations
Interest Rate System	Johansen	Rank Selection Criterion	Hypothesis
2-IR system: 1Y, 2Y	$\hat{r} = 1$	$\hat{r} = 1$	$r = 1$
3-IR system: 1Y, 2Y, 5Y	$\hat{r} = 1$	$\hat{r} = 1$	$r = 2$
4-IR system: 1Y, 2Y, 5Y, 7Y	$\hat{r} = 2$	$\hat{r} = 2$	$r = 3$
5-IR system: 1Y, 2Y, 5Y, 7Y, 10Y	$\hat{r} = 3$	$\hat{r} = 2$	$r = 4$

[1] Using the Bartlett-corrected trace statistic or the Bootstrap of Cavaliere et al. (2012), we obtain the same results as for Johansen's trace statistic.

both procedures estimate the cointegration rank to be one, the number implied by the expectations hypothesis. For the other interest rate systems, the estimated cointegration rank is lower than implied by the expectations hypothesis. In the 3-IR system, for instance, both procedures underestimate the cointegration rank implied by the theory (i.e. $r = 2$) by one (i.e. $\hat{r} = 1$). Empirical evidence for more than one common trend is also found by other researchers. Carstensen (2003) and Zhang (1993), for instance, find up to three common trends when interest rates of longer maturity our included. Giese (2008) find strong evidence for two common trends.

Zero-sum restriction. We test the null-hypothesis

$$H_0 : \mathbf{\Theta} = \begin{bmatrix} \theta_1 & \theta_2 & \dots & \theta_{q-1} \end{bmatrix}' = \mathbf{0_{(q-1)\times 1}}, \tag{18}$$

with $\theta_j = \sum_{i=1}^{q} \beta_{ij}$, $(j = 1, \dots, q-1)$ the sum of the coefficients of the jth cointegrating vector. Note that the zero-sum restriction implies the cointegration space to be perpendicular to the unit vector. Therefore, every basisvector of the cointegration space needs to be perpendicular to the unit vector.

We set the cointegration rank $r = q-1$, the number implied by the EHT, and estimate the cointegration space using Johansen's ML procedure or the sparse penalized ML procedure

resulting in an estimate $\widehat{\mathbf{\Theta}}$. To test the zero-sum restriction, we bootstrap the Wald test statistic $Q = \widehat{\mathbf{\Theta}}' \text{Cov}^{-1}(\widehat{\mathbf{\Theta}})\widehat{\mathbf{\Theta}}$. Details on the bootstrap procedure can be found in Appendix C.

Results are in Table 6. Johansen's procedure reports mixed evidence. The zero-sum restriction is rejected for the 3-, 4-, and 5-IR system (p-values < 0.05), but not for the 2-IR system (p-value > 0.05). Estimating the cointegrating vectors with a sparse estimator, the zero-sum restriction is not rejected (for all interest rate systems p-values > 0.05), confirming the EHT. Finally, note that many coefficients of the cointegrating vectors are estimated as exactly zero using the sparse penalized ML. This improves interpretability of the estimation results.

6.2 Forecasting industrial production in a large VECM. We consider a large VECM with $q = 24$ industrial production time series related to manufacturing, ranging from January 1972 until January 2014. We use an updated[3] version of the Stock and Watson (2002) database (source: Federal Reserve, United States). A short description of each time series can be found in Table 9, Appendix D. A stationarity test of all individual industrial production time series using the Augmented Dickey-Fuller test indicates that the time series are integrated of order one, making it appropriate to test for cointegration.

We use the Rank Selection Criterion from Section 4 to determine the cointegration rank since it performs much better than its alternatives in the high-dimensional simulation settings of Section 5. The Rank Selection Criterion indicates that the industrial production time series are cointegrated with 1 cointegration equation. The sparse method includes the time series `wood`, `prim metal`, `machinery`, `electrical` , `food` and `non-naics` in the cointegration equation as their associated coefficients are estimated as non-zero.

We compare the forecast performance of the Sparse penalized ML estimator (with Lasso penalty) to Johansen's ML. We estimate a VECM(1) model with one cointegration relation for the 24 industrial production time series. We take the order of the VECM to be one, as indicated by both the Bayesian Information Criterion and the Akaike Information Criterion.

[3]We extend the time range (until February 2014) and we add more industrial production time series.

Table 6: Testing the zero-sum restriction for each interest rate system using Johansen's ML and Sparse penalized ML with Lasso penalty. P-values are reported below every sum.

	Johansen ML				Sparse Lasso			
	Cointegrating vectors				Cointegrating vectors			
Variables	$\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_2$	$\hat{\boldsymbol{\beta}}_3$	$\hat{\boldsymbol{\beta}}_4$	$\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_2$	$\hat{\boldsymbol{\beta}}_3$	$\hat{\boldsymbol{\beta}}_4$
2-IR system								
1Y	1.00				1.00			
2Y	-1.01				-0.95			
sum	*-0.01*				*0.05*			
	$p = 0.91$				$p = 0.71$			
3-IR system								
1Y	1.00	1.00			1.00	1.00		
2Y	-2.41	-8.76			-1.52	0		
5Y	1.47	8.19			0.56	-0.87		
sum	*0.06*	*0.43*			*0.04*	*0.13*		
	$p < 0.01$				$p = 0.44$			
4-IR system								
1Y	1.00	1.00	1.00		1.00	1.00	1.00	
2Y	-19.78	-2.00	-6.72		-1.24	-0.97	-1.05	
5Y	48.29	0.43	4.30		0	0	0.06	
7Y	-29.78	0.64	1.79		0.27	0	0.04	
sum	*-0.27*	*0.07*	*0.37*		*0.03*	*0.03*	*0.05*	
	$p < 0.01$				$p = 0.62$			
5-IR system								
1Y	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2Y	-3.23	37.22	0.46	-4.11	-1.34	-1.07	-1.02	-0.80
5Y	1.31	-113.98	0.24	3.51	0	-0.11	0	0
7Y	3.61	86.16	-6.65	-3.70	0	0	0	0
10Y	-2.66	-9.67	5.02	3.61	0.35	0.19	0	-0.12
sum	*0.03*	*0.73*	*0.07*	*0.31*	*0.01*	*0.01*	*-0.02*	*0.08*
	$p < 0.01$				$p = 0.11$			

Note that we have included an intercept in the VECM of equation (1) since some of the industrial production time series exhibit a drift. Using a rolling window of 4 years (hence, $S = 48$), the VECM is re-estimated at each time point $t = S, \ldots, T-1$ and 1-step-ahead forecasts $\hat{\mathbf{y}}_{t+1} = (y_{t+1}^{(1)}, \ldots, y_{t+1}^{(24)})$ are computed. For each of the 24 time series ($i = 1 \ldots, q = 24$), we compute the Mean Absolute Forecast Error (MAFE).

$$\text{MAFE} = \frac{1}{T-S} \sum_{t=S}^{T-1} \mid \hat{y}_{t+1}^{(i)} - y_{t+1}^{(i)} \mid. \tag{19}$$

Table 7 reports the results for the two forecast methods.

Averaged across the 24 time series, the sparse estimation procedure achieves the best forecast performance. Its forecast performance is almost two times better than that of Johansen's ML (i.e. MAFE of 2.62 against 4.66). Also for 23 out of 24 industrial production time series, the MAFE of the Sparse Lasso is lower than the MAFE of Johansen's ML. A Diebold-Mariano test confirms that the forecast performance of the Sparse Lasso is significantly, at the 5% significance level, better than Johansen's ML, for 15 industrial production time series. In sum, we show that, in this high-dimensional application, sparsely estimating the cointegrating vector substantially improves the forecast performance compared to Johansen's approach.

7 Conclusion

In this paper, we discuss a sparse cointegration technique. Our simulation study shows that the sparse cointegration technique significantly outperforms Johansen's ML procedure, when the true cointegrating vectors are sparse or when the sample size is low compared to the number of time series. We use the Rank Selection Criterion of Bunea et al. (2011) to determine the cointegration rank. In high-dimensional simulation settings, the Rank Selection Criterion outperforms Johansen's trace statistic, the Bartlett-corrected trace statistic and the bootstrap procedure of Cavaliere et al. (2012).

Sparsity might be useful for several reasons. First, when the underlying structure of the cointegrating vectors is known to be sparse, a sparse cointegration technique allows to explic-

Table 7: Mean absolute forecast error (MAFE) for the $q = 24$ industrial production time series and the fwo forecast methods: Sparse penalized ML with Lasso penalty and Johansen's ML of the $q-$variate VECM with one cointegration relation. P-values of the Diebold-Mariano test are reported in the last column.

Time Series	Sparse Lasso	Johansen ML	$P-$value Diebold-Mariano test
TOT	1.58	1.74	0.54
NAICS	1.58	1.74	0.52
DURABLE	1.60	1.86	0.34
WOOD	3.27	6.56	< 0.01
NONMETAL	2.83	4.78	< 0.01
PRIM METAL	4.21	9.99	< 0.01
FABR METAL	1.94	2.23	0.41
MACHINERY	3.25	4.61	0.04
COMPUTER	1.14	0.99	0.47
ELECRICAL	2.95	5.03	< 0.01
MOTOR	4.40	11.51	< 0.01
AEROSPACE	3.31	5.22	0.01
FURNITURE	2.91	3.79	0.13
OTHER DURABLE	1.59	2.21	0.02
NONDURABLE	1.61	2.22	0.06
FOOD	1.62	3.61	< 0.01
TEXTILE	3.50	7.41	< 0.01
APPAREL	4.88	11.69	< 0.01
PAPER	2.38	6.02	< 0.01
PRINT	3.00	3.35	0.56
PETROLEUM	2.30	5.59	< 0.01
CHEMICAL	1.89	2.71	0.04
PLASTIC	2.45	3.26	0.04
NON-NAICS	2.78	3.70	0.09
Total	2.62	4.66	< 0.01

itly capture this sparseness. We illustrate this with the expectations hypothesis. Secondly, in high-dimensional settings with cointegrated time series, estimating the cointegrating vectors with a sparse estimator might improve estimation accuracy and/or forecast performance as illustrated in the industrial production application. Third, in over-parametrized settings, where the number of time series is larger than the sample size, traditional cointegration tests can not even be computed. The sparse estimator can be used in these settings.

There are several questions we did not address and which are left for future research. For instance, the models analyzed in this paper generally exclude deterministic terms (see e.g. Nielsen and Rahbek, 2000). An exception is the industrial production application where an intercept was included in the VECM. We also made abstraction of structural breaks. Allowing for structural breaks in the analysis is useful when analyzing macro-economic data (Johansen et al., 2000).

A natural extension of this study would be to implement structural analysis. Impulse-response functions, for instance, can be estimated using the sparse estimator. Confidence bound around the impulse-response functions are then obtained using a bootstrap procedure. Finally, similar ideas as introduced in this paper can be used to test for Granger Causality. Few studies consider Granger Causality in high-dimensional systems, an example is Jarocinski and Mackowiak (2013). An interesting path for future research is to use a sparse procedure to test for Granger Causality in high-dimensions.

Acknowledgments

Financial support from the FWO (Research Foundation Flanders) is gratefully acknowledged (FWO, contract number 11N9913N).

Appendix A: Simulation designs

Table 8: Low-dimensional ($T = 500, q = 4$) and high-dimensional ($T = 50, q = 11$) simulation designs.

Low-dimensional designs	$\boldsymbol{\beta}$	$\boldsymbol{\alpha}$	$\boldsymbol{\Gamma}_1$
Sparse $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}, r = 1$	$\boldsymbol{\beta}_1 = \begin{bmatrix} 1 \\ \mathbf{0}_{3\times 1} \end{bmatrix}$	$a\boldsymbol{\beta}_1$	$\boldsymbol{\Gamma}_1 = \gamma \boldsymbol{I}_q$
Sparse $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}, r = 2$	$\boldsymbol{\beta}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \mathbf{0}_{2\times 1} & \mathbf{0}_{2\times 1} \end{bmatrix}$	$a\boldsymbol{\beta}_2$	$\boldsymbol{\Gamma}_1 = \gamma \boldsymbol{I}_q$
Nonsparse $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}, r = 1$	$\boldsymbol{\beta}_3 = \begin{bmatrix} 1 \\ 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}$	$a\boldsymbol{\beta}_3$	$\boldsymbol{\Gamma}_1 = \gamma \boldsymbol{I}_q$
	with $a = -0.2, -0.4, \ldots, -0.8$, and $\gamma = 0.1$		
High-dimensional designs	$\boldsymbol{\beta}$	$\boldsymbol{\alpha}$	$\boldsymbol{\Gamma}_1$
Sparse $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}, r = 1$	$\boldsymbol{\beta}_4 = \begin{bmatrix} \mathbf{1}_{3\times 1} \\ \mathbf{0}_{8\times 1} \end{bmatrix}$	$a\boldsymbol{\beta}_4$	$\boldsymbol{\Gamma}_1 = \gamma \boldsymbol{I}_q$
Sparse $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}, r = 4$	$\boldsymbol{\beta}_5 = \begin{bmatrix} \mathbf{1}_{3\times 1} & \mathbf{0}_{3\times 1} & \mathbf{0}_{3\times 1} & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{3\times 1} & \mathbf{1}_{3\times 1} & \mathbf{0}_{3\times 1} & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{3\times 1} & \mathbf{0}_{3\times 1} & \mathbf{1}_{3\times 1} & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{2\times 1} & \mathbf{0}_{2\times 1} & \mathbf{0}_{2\times 1} & \mathbf{1}_{2\times 1} \end{bmatrix}$	$a\boldsymbol{\beta}_5$	$\boldsymbol{\Gamma}_1 = \gamma \boldsymbol{I}_q$
Nonsparse $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}, r = 1$	$\boldsymbol{\beta}_6 = \begin{bmatrix} \mathbf{1}_{3\times 1} \\ \mathbf{0.1}_{8\times 1} \end{bmatrix}$	$a\boldsymbol{\beta}_6$	$\boldsymbol{\Gamma}_1 = \gamma \boldsymbol{I}_q$
	with $a = -0.2, -0.4, \ldots, -0.8$, and $\gamma = 0.4$		

Appendix B: Time-series cross-validation

We select the tuning parameters according to a time series cross-validation approach (Hyndman, 2014). Denote the response by $\mathbf{z}_t$. For the penalized multivariate regression in equation (8), $\mathbf{z}_t = \boldsymbol{\Delta}\mathbf{y}_t - \boldsymbol{\Pi}\mathbf{y}_{t-1}$. For the penalized reduced rank regression in equation (9), $\mathbf{z}_t = \boldsymbol{\Delta}\mathbf{y}_t - \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i \boldsymbol{\Delta}\mathbf{y}_{t-i}$.

1. For $t = S, \ldots, T-1$ (with S such that 80% of the data is included in the first calibration sample), repeat:
   (a) For a grid of tuning parameters, fit the model to the data $\mathbf{z}_1, \ldots, \mathbf{z}_t$.
   (b) Compute the one-step-ahead forecast error $\mathbf{e}_{t+1} = \mathbf{z}_{t+1} - \hat{\mathbf{z}}_{t+1}$
2. Select the value of the tuning parameter that minimizes the mean squared forecast error

$$\text{MSFE} = \frac{1}{T-S}\frac{1}{q}\sum_{t=S}^{T-1}\sum_{i=1}^{q}\left(\frac{e_{t+1}^{(i)}}{\hat{\sigma}^{(i)}}\right)^2,$$

   with $e_t^{(q)}$ the q^{th} component of the multivariate time series at time t and $\hat{\sigma}^{(q)}$ the standard deviation of the time series $z_t^{(q)}$.

Appendix C: Bootstrap procedure for testing the zero-sum restriction

To test the zero-sum restriction, we use the following bootstrap procedure (see Cavaliere et al., 2012).

1. Take the cointegrating vector under the null hypothesis, $\boldsymbol{\beta}^{H_0}$, see equation (17). Given $\boldsymbol{\beta}^{H_0}$, use the Sparse penalized ML algorithm (or Johansen's approach) to estimate $\hat{\boldsymbol{\alpha}}^{H_0}, \widehat{\boldsymbol{\Gamma}}_1^{H_0}, \ldots, \widehat{\boldsymbol{\Gamma}}_{p-1}^{H_0}$, together with the corresponding centered residuals $\hat{\boldsymbol{\varepsilon}}_t$.
2. Construct the bootstrap sample recursively from

$$\boldsymbol{\Delta}\mathbf{y}_t^{H_0*} = \hat{\boldsymbol{\alpha}}^{H_0}\boldsymbol{\beta}^{H_0\prime}\mathbf{y}_{t-1}^* + \sum_{i=1}^{p-1}\widehat{\boldsymbol{\Gamma}}_i^{H_0}\boldsymbol{\Delta}\mathbf{y}_{t-i}^* + \boldsymbol{\varepsilon}_t^*,$$

with starting values $\mathbf{y}_t^* = \mathbf{y}_j, j = 1-p, \ldots, 0$ and with bootstrap errors $\boldsymbol{\varepsilon}_t^*$ obtained using a residual bootstrap such that $\boldsymbol{\varepsilon}_t^* = \hat{\boldsymbol{\varepsilon}}_{\mathcal{U}_t}$ with $\mathcal{U}_t, t = 1, \ldots, T$ an i.i.d. sequence of discrete uniform distributions on $\{1, \ldots, T\}$.

3. Apply the Sparse penalized ML algorithm (or Johansen's approach) to the bootstrap sample $\mathbf{y}_t^{H_0*}$.

4. Construct the bootstrap estimates $\widehat{\boldsymbol{\Theta}}^{*\prime} = \begin{bmatrix} \hat{\theta}_1^* & \hat{\theta}_2^* & \ldots & \hat{\theta}_{q-1}^* \end{bmatrix}'$, with $\hat{\theta}_j^* = \sum_{i=1}^{q} \hat{\beta}_{ij}^*$.

5. Compute the bootstrap statistic $Q^* = \widehat{\boldsymbol{\Theta}}^{*\prime} \mathrm{Cov}^{-1}(\widehat{\boldsymbol{\Theta}}^*) \widehat{\boldsymbol{\Theta}}^*$.

6. Check if $B^{-1} \sum_{b=1}^{B} \mathbf{1}(Q_b^* > Q)$ - with $Q_b^*, b = 1, \ldots, B$ B independent bootstrap statistics - exceeds a fixed significance level η. If so, the null hypothesis H_0 is not rejected.

Appendix D: Industrial Production Time Series

Table 9: Industrial production time series. Source: Federal Reserve, United States

Variable	Description
TOT	Total manufacturing
NAICS	NAICS industry manufacturing
DURABLE	Durable manufacturing
WOOD	Wood production
NONMETAL	Nonmetallic mineral production
PRIM METAL	Primary metal
FABR METAL	Fabricated metal
MACHINERY	Machinery
COMPUTER	Computer and Electronic product
ELECRICAL	Electrical equipment, appliance and component
MOTOR	Motor vehicles and parts
AEROSPACE	Aerospace and other miscellaneous transportation
FURNITURE	Furniture and related products
OTHER DURABLE	Miscellaneous durable manufacturing
NONDURABLE	Nondurable manufacturing
FOOD	Food, beverage and tobacco
TEXTILE	Textile and production mills
APPAREL	Nondurables, apparel and leather goods
PAPER	Paper
PRINT	Printing and related support activities
PETROLEUM	Petroleum and coal products
CHEMICAL	Chemical
PLASTIC	Plastics and rubber products
NON-NAICS	non-NAICS industry manufacturing

References

Banbura, M.; Giannone, D. and Reichlin, L. (2010), "Large Bayesian vector auto regressions," *Journal of Applied Econometrics*, 25, 71–92.

Bernardini, E. and Cubadda, G. (2014), "Macroeconomic forecasting and structural analysis through regularized reduced-rank regression," *International Journal of Forecasting*, In Press, Available online 1 February 2014.

Breitung, J. and Cubadda, G. (2011), "Testing for cointegration in high-dimensional systems," *CEIS Working Paper*.

Bunea, F.; She, Y. and Wegkamp, M. (2011), "Optimal selection of reduced rank estimators of high-dimensional matrices," *The Annals of Statistics*, 39, 1282–1309.

Caner, M. and Zhang, H. (2014), "Adaptive elastic net for generalized methods of moments," *Journal of Business and Economic Statistics*, 32, 30–47.

Carriero, A.; Kapetanios, G. and Marcellino, M. (2011), "Forecasting large datasets with Bayesian reduced rank multivariate models," *Journal of Applied Econometrics*, 26, 735–761.

Carstensen, K. (2003), "Nonstationary term premia and cointegration of the term structure," *Economic Letters*, 80, 409–413.

Cavaliere, G.; Rahbek, A. and Taylor, A. R. (2012), "Bootstrap determination of the co-integration rank in vector autoregressive models," *Econometrica*, 80, 1721–1740.

Chen, L. and Huang, J. (2012), "Sparse reduced-rank regression for simultaneous dimension reduction and variable selection," *Journal of the American Statistical Association*, 107, 1533–1545.

Clements, M. and Galvao, A. (2008), "Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the United States," *Journal of Business and Economic Statistics*, 26, 546–554.

Doornik, J. (1998), "Approximations to the asymptotic distribution of cointegration tests," *Journal of Economic Surveys*, 12, 573–593.

Engle, R. and Granger, C. (1987), "Cointegration and error correction: representation, estimation, and testing," *Econometrica*, 55, 251–276.

Engsted, T. and Tanggaard, C. (1994), "Cointegration and the US term structure," *Journal of Banking and Finance*, 18, 167–181.

Fan, J.; Lv, J. and Qi, L. (2011), "Sparse high-dimensional models in economics," *Annual Review of Economics*, 3, 291–317.

Friedman, J.; Hastie, T. and Tibshirani, R. (2008), "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, 9, 432–441.

Giese, J. (2008), "Level, slope, curvature: Characterising the yield curve in a cointegrated VAR model," *Economics*, 2, No. 2008–28.

Hall, A.; Anderson, H. and Granger, C. (1992), "A cointegration analysis of treasury bill yields,"

Review of Economics and Statistics, 74, 116–126.

Hamilton, J. (1991), *Time Series Analysis*, Princeton University Press.

Hoerl, A. and Kennard, R. (1970), "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, 12, 55–67.

Hyndman, R. (2014), *forecast: Forecasting functions for time series and linear models*, R package version 5.2.

Jarocinski, M. and Mackowiak, B. (2013), "Granger-causal-priority and choice of variables in vector autoregressions," *ECB Working Paper Series*, No. 1600.

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

— (1996), *Likelihood-based inference in cointegrated vector autoregressive models*, Oxford: Oxford University Press.

— (2002), "A small sample correction of the test for cointegration rank in the vector autoregressive model," *Econometrica*, 70, 1929–1961.

Johansen, S.; Mosconi, R. and Nielsen, B. (2000), "Cointegration analysis in the presence of structural breaks in the deterministic trend," *Econometrics Journal*, 3, 216–249.

Juselius, K. (2006), *The cointegrated VAR model*, Oxford: Oxford University Press.

Koop, G.; Strachan, R.; Van Dijk, H. and Villani, M. (2006), *Bayesian approaches to cointegration. In: The Palgrave Handbook of Theoretical Econometrics*, Palgrave Macmillan.

Korobilis, D. (2013), "VAR forecasting using Bayesian variable selection," *Journal of Applied Econometrics*, 28, 204–230.

Lanne, M. (2000), "Near unit roots, cointegration, and the term structure of interest rates," *Journal of Applied Econometrics*, 15, 513–529.

Lasak, K. and Velasco, C. (2014), "Fractional cointegration rank estimation," *Journal of Business and Economic Statistics*, Accepted manuscript.

Lissitz, R.; Schonemann, P. and Lingoes, J. (1976), "A solution to the weighted Procrustes problem in which the transformation is in agreement with the loss function," *Psychometrika*, 41, 547–550.

Lutkepohl, H. (2007), *New introduction to multiple time series analysis*, Springer-Verlag.

Nielsen, B. (2004), "On the distribution of tests of cointegration," *Econometric Reviews*, 23, 1–23.

Nielsen, B. and Rahbek, A. (2000), "Similarity issues in cointegration models," *Oxford Bulletin of Economics and Statistics*, 62, 5–22.

Osterwald-Lenum, M. (1992), "A note with quantiles of the asymptotic distribution of the maximum likelihood cointegration rank test statistics," *Oxford Bulletin of Economics and Statistics*, 54, 461–472.

Phillips, P. and Ouliaris, S. (1990), "Asymptotic properties of residual based tests for cointegration," *Econometrica*, 58, 165–193.

Rothman, A.; Levina, E. and Zhu, J. (2010), "Sparse multivariate regression with covariance esti-

mation," *Journal of Computational and Graphical Statistics*, 19, 947–962.

Shea, G. (1992), "Benchmarking the expectations hypothesis of the interest-rate term structure: An analysis of cointegration vectors," *Journal of Business and Economic Statistics*, 10, 347–366.

Stock, J. and Watson, M. (2002), "Macroeconomic forecasting using diffusion indexes," *Journal of Business and Economic Statistics*, 20, 147–162.

Strachan, R. (2003), "Valid Bayesian estimation of the cointegrating error correction model," *Journal of Business and Economic Statistics*, 21, 185–195.

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

Zhang, H. (1993), "Treasury yield curves and cointegration," *Applied Economics*, 25, 361–367.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.

FACULTY OF ECONOMICS AND BUSINESS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be