International Journal of Forecasting 32 (2016) 1256–1267

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Forecasting using sparse cointegration

Ines Wilms *, Christophe Croux

Faculty of Economics and Business, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium

ARTICLE INFO

Keywords:
Lasso
Reduced rank regression
Sparse estimation
Time series forecasting
Vector error correction model

ABSTRACT

This paper proposes a sparse cointegration method. Cointegration analysis is used to estimate the long-run equilibrium relationships between several time series, with the coefficients of these long-run equilibrium relationships being the cointegrating vectors. We provide a sparse estimator of the cointegrating vectors, where sparse estimation means that some elements of the cointegrating vectors are estimated to be exactly zero. The sparse estimator is applicable in high-dimensional settings, where the time series is short compared to the number of time series. Our method achieves better estimation and forecast accuracy than the traditional Johansen method in sparse and/or high-dimensional settings. We use the sparse method for interest rate growth forecasting and consumption growth forecasting. The sparse cointegration method leads to important forecast accuracy gains relative to the Johansen method.

1. Introduction

High-dimensional data sets containing thousands of time series are commonly available and can be accessed at a reasonable cost (Fan, Lv, & Qi, 2011; Stock & Watson, 2002). Recently, there has been a considerable amount of work on exploiting the large amount of information contained in these data sets for forecasting purposes. To handle the dimensionality, various large time series models, containing large numbers of time series relative to the time series length, have been considered. Common approaches include factor models (e.g., Stock & Watson, 2002), Bayesian vector autoregressive (VAR) models (e.g., Banbura, Giannone, & Reichlin, 2010), and reduced-rank VAR models (e.g., Carriero, Kapetanios, & Marcellino, 2011 and Bernardini & Cubadda, 2015), among others. Typically, though, these authors have not accounted for cointegration. Instead, either the time series are transformed in order to achieve stationarity (Bernardini & Cubadda, 2015), or the (non-)stationarity is accounted for in the prior distribution of the autoregressive parameters (Banbura et al., 2010). In cointegration analysis, we estimate long-run equilibrium relationships between several time series, often as implied by economic theory.

This paper develops a cointegration method for high-dimensional time series. The vector error correction model (VECM; e.g., Lütkepohl, 2007) is used to estimate and test for the cointegration relationships. Various cointegration tests exist (e.g., Engle & Granger, 1987 and Phillips & Ouliaris, 1990), with the cointegration test of Johansen (1988) being the most popular. However, Johansen's maximum likelihood approach has various limitations. In a high-dimensional setting, where the number of time series is large compared to the length of the time series, the estimation imprecision will be large. Johansen's approach is based on the estimation of a VAR model and a canonical correlation analysis. One drawback of the VAR is that the number of parameters that it uses increases quadratically with the number of time series included. As a consequence, the regression parameters will be estimated inaccurately if only limited numbers of time points are available. When

* Corresponding author.
E-mail addresses: Ines.Wilms@kuleuven.be (I. Wilms), Christophe.Croux@kuleuven.be (C. Croux).

http://dx.doi.org/10.1016/j.ijforecast.2016.04.005
0169-2070/

the number of time series exceeds the time series length, Johansen's approach cannot even be applied.

We introduce a penalized maximum likelihood (PML) approach that is designed for estimating the cointegrating vectors in a sparse way, i.e., with some of its components estimated as exactly zero. Sparse estimators have been shown to perform well in various fields, such as economics (e.g., Fan et al., 2011), macroeconomics (e.g., Korobilis, 2013; Liao & Phillips, 2015), finance (e.g., Zhou, Nakajima, & West, 2014), and biostatistics (e.g., Friedman, 2012). Sparse cointegration methods are useful for several reasons. First, sparsity facilitates model interpretation, since only limited numbers of time series, those corresponding to the non-zero coefficients, enter the estimated long-run equilibrium relationships. Second, sparsity improves the forecast performance through a variance reduction. Third, unlike Johansen's maximum likelihood approach, the sparse approach can still be applied when the number of time series exceeds the time series length.

We show in a simulation study that the sparse cointegration method outperforms Johansen's method significantly when the cointegrating vectors are sparse or when the number of time series is large compared to the time series length. Furthermore, we evaluate the forecast performance of the proposed sparse cointegration method on two data sets. We show that important gains in forecast accuracy can be obtained by accounting for cointegration and by estimating the cointegrating vectors sparsely.

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, She, & Wegkamp, 2011) for determining the cointegration rank. Section 5 presents the results of a simulation study. Section 6 discusses two forecasting examples: first we forecast interest rate growth, then we forecast consumption growth. Finally, Section 7 concludes.

2. Penalized maximum likelihood

Let $\boldsymbol{y}_t$ be a q-dimensional multivariate time series. We assume that the vector process $\boldsymbol{y}_t$ is integrated of order one $I(1)$, meaning that its first difference is stationary. Note that $\boldsymbol{y}_t$ can be $I(1)$ even if some of its components are stationary (Johansen, 1991, Ch. 5). Furthermore, we assume that $\boldsymbol{y}_t$ follows a vector autoregressive model of order p, denoted VAR(p). Any pth order VAR can be rewritten in a vector error correction (VECM) representation (Hamilton, 1991) as follows:

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

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 $N_q(\boldsymbol{0}, \boldsymbol{\Sigma})$.

If we can express $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$, with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ being $q \times r$ matrices of full column rank r, with $0 < r < q$, then the linear combinations given by $\boldsymbol{\beta}'\boldsymbol{y}_t$ are stationary and $\boldsymbol{y}_t$ 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 by penalized maximum likelihood (PML). It is convenient to rewrite Eq. (1) in matrix notation:

$$\Delta \boldsymbol{Y} = \Delta \boldsymbol{Y}_L \boldsymbol{\Gamma} + \boldsymbol{Y} \boldsymbol{\Pi}' + \boldsymbol{E}, \tag{2}$$

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

Consider the penalized negative log-likelihood

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Gamma}, \boldsymbol{\Pi}, \boldsymbol{\Omega}) = \frac{1}{T} \operatorname{tr}\Big(& (\Delta \boldsymbol{Y} - \Delta \boldsymbol{Y}_L \boldsymbol{\Gamma} - \boldsymbol{Y} \boldsymbol{\Pi}') \\ & \times \boldsymbol{\Omega} (\Delta \boldsymbol{Y} - \Delta \boldsymbol{Y}_L \boldsymbol{\Gamma} - \boldsymbol{Y} \boldsymbol{\Pi}')' \Big) - \log |\boldsymbol{\Omega}| \\ & + \lambda_1 P_1(\boldsymbol{\beta}) + \lambda_2 P_2(\boldsymbol{\Gamma}) + \lambda_3 P_3(\boldsymbol{\Omega}), \end{aligned} \tag{3}$$

with $\operatorname{tr}(\cdot)$ denoting the trace, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, and P_1, P_2 and P_3 being three penalty functions. We use L_1 penalization (see Tibshirani, 1996, in reference to the lasso) on the cointegrating vectors $\boldsymbol{\beta}$:

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

By adding the L_1 penalty to the objective function in Eq. (3), we obtain a sparse solution: some elements of $\boldsymbol{\beta}$ are estimated to be exactly zero. We use L_1 penalization on the short-run effects $\boldsymbol{\Gamma}$ and the off-diagonal elements of the inverse error covariance matrix $\boldsymbol{\Omega}$ similarly.

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

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

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $q \times r$ matrices of full column rank r. The matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are not defined uniquely. For identifiability purposes, we impose the normalization conditions $\boldsymbol{\alpha}'\boldsymbol{\Omega}\boldsymbol{\alpha} = \boldsymbol{I}_r$. For the unpenalized case ($\lambda_1 = 0$, $\lambda_2 = 0$ and $\lambda_3 = 0$), the objective function in Eq. (3) boils down to that introduced by Johansen (1988). The unpenalized case can be solved using either the closed-form expressions of Johansen (1988) or the iterative algorithm described below.

3. Algorithm

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

3.0.1. Solving for $\boldsymbol{\Pi}$ conditional on $\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}$

When $\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}$ are fixed, the minimization problem in Eq. (3) with $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ is equivalent to

$$\begin{aligned} (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) | \boldsymbol{\Gamma}, \boldsymbol{\Omega} = \underset{\boldsymbol{\alpha}, \boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{T} \operatorname{tr}\Big(& (\Delta \boldsymbol{Y} - \Delta \boldsymbol{Y}_L \boldsymbol{\Gamma} - \boldsymbol{Y} \boldsymbol{\beta}\boldsymbol{\alpha}') \\ & \times \boldsymbol{\Omega} (\Delta \boldsymbol{Y} - \Delta \boldsymbol{Y}_L \boldsymbol{\Gamma} - \boldsymbol{Y} \boldsymbol{\beta}\boldsymbol{\alpha}')' \Big) + \lambda_1 P_1(\boldsymbol{\beta}), \end{aligned} \tag{5}$$

which boils down to a penalized reduced rank regression (Chen & Huang, 2012). We begin by estimating $\boldsymbol{\alpha}$ conditional on $\boldsymbol{\beta}$, then estimate $\boldsymbol{\beta}$ conditional on $\boldsymbol{\alpha}$.

For a fixed $\boldsymbol{\beta}$, the minimization problem in Eq. (5) reduces to

$$\hat{\boldsymbol{\alpha}}|\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\beta} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{T}\operatorname{tr}\Big((\Delta\boldsymbol{Y} - \Delta\boldsymbol{Y}_L\boldsymbol{\Gamma} - \boldsymbol{Y}\boldsymbol{\beta}\boldsymbol{\alpha}') \times \boldsymbol{\Omega}(\Delta\boldsymbol{Y} - \Delta\boldsymbol{Y}_L\boldsymbol{\Gamma} - \boldsymbol{Y}\boldsymbol{\beta}\boldsymbol{\alpha}')'\Big),$$

subject to $\boldsymbol{\alpha}'\boldsymbol{\Omega}\boldsymbol{\alpha} = \boldsymbol{I}_r$, which is a weighted Procrustes problem (Lissitz, Schonemann, & Lingoes, 1976). This weighted Procrustes problem for $\boldsymbol{\alpha}$ can be seen as an unweighted Procrustes problem for $\boldsymbol{\alpha}^\star = \boldsymbol{\Omega}^{1/2}\boldsymbol{\alpha}$. 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

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

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

For a fixed $\boldsymbol{\alpha}$, the minimization problem in Eq. (5) reduces to

$$\hat{\boldsymbol{\beta}}|\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\alpha} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{T}\operatorname{tr}\Big((\Delta\boldsymbol{Y} - \Delta\boldsymbol{Y}_L\boldsymbol{\Gamma} - \boldsymbol{Y}\boldsymbol{\beta}\boldsymbol{\alpha}') \times \boldsymbol{\Omega}(\Delta\boldsymbol{Y} - \Delta\boldsymbol{Y}_L\boldsymbol{\Gamma} - \boldsymbol{Y}\boldsymbol{\beta}\boldsymbol{\alpha}')'\Big) + \lambda_1 P_1(\boldsymbol{\beta}). \quad (6)$$

Since $\boldsymbol{\alpha}^{\star\prime}\boldsymbol{\alpha}^\star = \boldsymbol{I}_r$, 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 $\tilde{\boldsymbol{Y}} = \Delta\boldsymbol{Y} - \Delta\boldsymbol{Y}_L\boldsymbol{\Gamma}$,

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

where $\|\cdot\|$ denotes the Frobenius norm for a matrix. Since the second term on the left-hand-side does not involve $\boldsymbol{\beta}$, the minimization problem reduces to

$$\hat{\boldsymbol{\beta}}|\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\alpha} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{T}\operatorname{tr}\Big((\tilde{\boldsymbol{Y}}\boldsymbol{\Omega}^{1/2}\boldsymbol{\alpha}^\star - \boldsymbol{Y}\boldsymbol{\beta}) \times (\tilde{\boldsymbol{Y}}\boldsymbol{\Omega}^{1/2}\boldsymbol{\alpha}^\star - \boldsymbol{Y}\boldsymbol{\beta})'\Big) + \lambda_1 P_1(\boldsymbol{\beta}), \quad (7)$$

which is a penalized multivariate least squares regression of $\tilde{\boldsymbol{Y}}\boldsymbol{\Omega}^{1/2}\boldsymbol{\alpha}^\star$ on $\boldsymbol{Y}$.

3.0.2. Solving for $\boldsymbol{\Gamma}$ conditional on $\boldsymbol{\Pi}$ and $\boldsymbol{\Omega}$

When $\boldsymbol{\Pi}$ and $\boldsymbol{\Omega}$ are fixed, the minimization problem in Eq. (3) is a penalized multivariate regression of $\big(\Delta\boldsymbol{Y} - \boldsymbol{Y}\boldsymbol{\Pi}'\big)$ on $\Delta\boldsymbol{Y}_L$ (Rothman, Levina, & Zhu, 2010).

3.0.3. Solving for $\boldsymbol{\Omega}$ conditional on $\boldsymbol{\Gamma}$ and $\boldsymbol{\Pi}$

When $\boldsymbol{\Gamma}$ and $\boldsymbol{\Pi}$ are fixed, the minimization problem in Eq. (3) corresponds to penalized covariance estimation (Friedman, Hastie, & Tibshirani, 2008).

3.1. Convergence criterion

We iterate our solving of the minimization problems described above until the relative change in the value of the objective function, i.e., the penalized log-likelihood in Eq. (3), in two successive iterations[1] is smaller than a prespecified tolerance level ϵ, chosen to be $\epsilon = 10^{-2}$. Although there is no proof of convergence of the algorithm, we have observed it empirically in all real data examples and all simulation runs. For a data set (generated as in the simulation study in Section 5) consisting of $q = 4$ time series, each of length $T = 500$, an average of three iterations were needed for convergence, while four iterations were needed for convergence on average with $q = 11$ and $T = 50$.

3.2. Selection of tuning parameters

Tuning parameters are selected at each step of the iterative algorithm. 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, based on a time series cross-validation approach (Hyndman, 2014), see Appendix A. 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). As a default, we use a grid of one hundred λ_1 values, five λ_2 values and five λ_3 values.

3.3. Starting values

Starting values for $\boldsymbol{\Omega}$, $\boldsymbol{\Gamma}$ and $\boldsymbol{\beta}$ are required. We take the identity matrices for $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}_k$, $k = 1, \ldots, p-1$. For $\boldsymbol{\beta}$, we take the first r eigenvectors of the matrix $\widehat{\Sigma}_{YY}^{-1}\widehat{\Sigma}_{Y\Delta Y}\widehat{\Sigma}_{\Delta Y\Delta Y}^{-1}\widehat{\Sigma}_{\Delta YY}$, where we take $\widehat{\Sigma}_{YY}$ and $\widehat{\Sigma}_{\Delta Y\Delta Y}$ to be diagonal and $\widehat{\Sigma}_{Y\Delta Y} = \widehat{\Sigma}_{\Delta YY}'$ to be the sample covariance matrix between $\boldsymbol{Y}$ and $\Delta\boldsymbol{Y}$.

We performed several numerical experiments to investigate the robustness of the outcome of the algorithm to the choice of starting values. The choice of starting values is unimportant in low-dimensional settings, but more important in high-dimensional settings. Note that the starting values should exist and be easy to compute in all settings, which holds for our proposal.

3.4. Computation time

All computations are carried out in R version 3.2.1, and the code of the algorithm is available on the homepage of the first author (http://feb.kuleuven.be/public/n12066/SparseCointegration). The PML estimator is quite quick to

[1] One iteration includes one cycle of estimating $\boldsymbol{\Pi}|\boldsymbol{\Gamma}, \boldsymbol{\Omega}$; $\boldsymbol{\Gamma}|\boldsymbol{\Pi}, \boldsymbol{\Omega}$; and $\boldsymbol{\Omega}|\boldsymbol{\Gamma}, \boldsymbol{\Pi}$.

compute: on an Intel Core i7-3720QM @ 2.60 GHz machine, the computation takes eight seconds on average for a data set consisting of $q = 4$ time series, each of length $T = 500$, and four seconds on average for a data set with $q = 11$ and $T = 50$. This computation time includes the cross-validation for the selection of tuning parameters.

4. Determination of cointegration rank

In small, finite samples, the asymptotic distribution of Johansen's trace statistic, used to determine the cointegration rank, might be a poor approximation of the true distribution, resulting in substantial size and power distortions (e.g., Johansen, 2002 and Nielsen, 2004). We determine the cointegration rank r using an iterative procedure based on the rank selection criterion (RSC) of Bunea et al. (2011). We start with an initial value of the cointegration rank of $r_{\text{start}} = q$.

For this initial value, we obtain $\widehat{\mathbf{\Gamma}}$ using the algorithm in Section 3. Next, 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 $\widetilde{\Delta \boldsymbol{Y}}' \boldsymbol{P} \widetilde{\Delta \boldsymbol{Y}}$ that exceed the threshold μ:

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

with $\widetilde{\Delta \boldsymbol{Y}} = \Delta \boldsymbol{Y} - \Delta \boldsymbol{Y}_L \widehat{\mathbf{\Gamma}}$, and $\boldsymbol{P} = \boldsymbol{Y}(\boldsymbol{Y}'\boldsymbol{Y})^{-}\boldsymbol{Y}'$ being the projection matrix onto the column space of $\boldsymbol{Y}$. Note that $(\boldsymbol{Y}'\boldsymbol{Y})^{-}$ denotes the Moore–Penrose inverse of the matrix $(\boldsymbol{Y}'\boldsymbol{Y})$. Following the recommendation of Bunea et al. (2011), the threshold is set equal to $\mu = 2S^2(q + l)$, under the assumption that $l < T$, with $l = \text{rank}(\boldsymbol{Y})$ and

$$S^2 = \frac{\|\widetilde{\Delta \boldsymbol{Y}} - \boldsymbol{P}\widetilde{\Delta \boldsymbol{Y}}\|^2}{Tq - lq}.$$

We repeat the above procedure using the new value of $\hat{r}$ until the estimated cointegration rank does not change in two successive iterations.

The rank selection criterion consistently estimates 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.

5. Simulation study

We conduct a simulation study to evaluate the performance of the PML estimator. The data generating process (revised from Cavaliere, Rahbek, & Taylor, 2012) is the following VECM:

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

where the error terms $\boldsymbol{e}_t$ follow a $N_q(\mathbf{0}, \mathbf{\Sigma})$ distribution. We set $\boldsymbol{y}_0 = \Delta \boldsymbol{y}_0 = \mathbf{0}$. All simulated models satisfy the assumptions of the VECM described in Section 2.

We compare the out-of-sample forecast accuracies of the PML estimator and the ML estimator of Johansen (1988), and find that the former performs significantly better than the latter in sparse and/or high-dimensional settings. In addition, we also compare their estimation accuracies and investigate the performance of the rank selection criterion in selecting the true cointegration rank correctly.

5.1. Simulation designs

Two different simulation designs are considered: (i) low-dimensional ($T = 500, q = 4$), and (ii) high-dimensional with moderate time series length ($T = 50$, $q = 11$).[2] We consider both sparse and non-sparse settings, and report on selected representative cases below. Full details of each selected setting are given in Table 1.

5.1.1. Low-dimensional designs

The true cointegrating vectors and the short-run effects are sparse in the first two simulation settings, and non-sparse in the third. The cointegration ranks are equal to $r = 1$, $r = 2$ and $r = 1$, respectively. While $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ belong to different spaces in the first and third settings, they belong to the same space in setting two. Furthermore, the error terms of the VECM are uncorrelated in settings one and three, but correlated in setting two.

5.1.2. High-dimensional designs

The true cointegrating vectors and the short-run effects are sparse in the first two simulation settings and non-sparse in the third. The cointegration ranks are equal to $r = 1$, $r = 4$ and $r = 1$, respectively. The choices for the relationship between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and the error terms are similar to those of the low-dimensional designs.

5.2. Estimation accuracy

We evaluate the estimation accuracy by computing the angle $\theta^{(m)}(\hat{\boldsymbol{\beta}}^{(m)}, \boldsymbol{\beta})$ between the estimated cointegration space and the true cointegration space for each simulation run m, with $m = 1, \ldots, M = 500$.[3] The average angle is then given by

$$\theta(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \frac{1}{M}\sum_{m=1}^{M} \theta^{(m)}(\hat{\boldsymbol{\beta}}^{(m)}, \boldsymbol{\beta}). \tag{8}$$

The value of the angle varies from zero (for identical subspaces) to $\pi/2$ (for orthogonal subspaces).

5.2.1. Results

Simulation results on the accuracy of the estimated cointegration space are given in Table 2, which reports the average angle (averaged across simulation runs) between

[2] The largest number for which the critical values of Johansen's trace statistic are tabulated by Johansen (1996, Ch. 15) is $q = 11$ time series.

[3] The angle $\theta^{(m)}(\hat{\boldsymbol{\beta}}^{(m)}, \boldsymbol{\beta})$ is computed as follows (see e.g. Anderson, 1958). First, compute the QR-decompositions $\widehat{\boldsymbol{\beta}}^{m} = \boldsymbol{Q}_{\widehat{\boldsymbol{\beta}}^m}\boldsymbol{R}_{\widehat{\boldsymbol{\beta}}^m}$ and $\boldsymbol{\beta} = \boldsymbol{Q}_{\boldsymbol{\beta}}\boldsymbol{R}_{\boldsymbol{\beta}}$. Next, compute the singular value decomposition of $\boldsymbol{Q}_{\widehat{\boldsymbol{\beta}}^m}^{T}\boldsymbol{Q}_{\boldsymbol{\beta}} = \boldsymbol{U}\boldsymbol{C}\boldsymbol{V}^T$. The matrix $\boldsymbol{C}$ is diagonal, with elements $c_1 \geq \cdots \geq c_r$, and the minimum angle is given by $\theta^{(m)}(\hat{\boldsymbol{\beta}}^{(m)}, \boldsymbol{\beta}) = \cos^{-1}(c_1)$.

Table 1
Low-dimensional ($T = 500$, $q = 4$) and high-dimensional ($T = 50$, $q = 11$) simulation designs.

Low-dimensional designs	$\boldsymbol{\beta}$	$\boldsymbol{\alpha}$	$\boldsymbol{\Gamma}_1$	$\boldsymbol{\Sigma}$
Sparse $r = 1$	$\begin{bmatrix} 1 \\ \mathbf{0}_{3\times 1} \end{bmatrix}$	$a \cdot \begin{bmatrix} \mathbf{1}_{2\times 1} \\ \mathbf{0}_{2\times 1} \end{bmatrix}$	$\gamma \boldsymbol{I}_q$	$\boldsymbol{I}_q$
Sparse $r = 2$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \mathbf{0}_{2\times 1} & \mathbf{0}_{2\times 1} \end{bmatrix}$	$a\boldsymbol{\beta}$	$\gamma \boldsymbol{I}_q$	$\Sigma_{ij} = 0.2^{\lvert i-j \rvert}$
Non-sparse $r = 1$	$\begin{bmatrix} 1 \\ \mathbf{0.1}_{3\times 1} \end{bmatrix}$	$a \cdot \begin{bmatrix} \mathbf{1}_{2\times 1} \\ \mathbf{0.1}_{2\times 1} \end{bmatrix}$	$\boldsymbol{\Gamma}_{1,ij} = \begin{cases} \gamma & \text{if } j = i \\ \gamma \cdot 10^{-4} & \text{if } j \neq i \end{cases}$	$\gamma \boldsymbol{I}_q$
	with $a = -0.2, -0.4, \ldots, -0.8$, and $\gamma = 0.1$			
High-dimensional designs	$\boldsymbol{\beta}$	$\boldsymbol{\alpha}$	$\boldsymbol{\Gamma}_1$	$\boldsymbol{\Sigma}$
Sparse $r = 1$	$\begin{bmatrix} \mathbf{1}_{3\times 1} \\ \mathbf{0}_{8\times 1} \end{bmatrix}$	$a \cdot \begin{bmatrix} \mathbf{1}_{6\times 1} \\ \mathbf{0}_{5\times 1} \end{bmatrix}$	$\gamma \boldsymbol{I}_q$	$\boldsymbol{I}_q$
Sparse $r = 4$	$\begin{bmatrix} \mathbf{1}_{3\times 1} & \mathbf{0}_{3\times 1} & \mathbf{0}_{3\times 1} & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{3\times 1} & \mathbf{1}_{3\times 1} & \mathbf{0}_{3\times 1} & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{3\times 1} & \mathbf{0}_{3\times 1} & \mathbf{1}_{3\times 1} & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{2\times 1} & \mathbf{0}_{2\times 1} & \mathbf{0}_{2\times 1} & \mathbf{1}_{2\times 1} \end{bmatrix}$	$a\boldsymbol{\beta}$	$\gamma \boldsymbol{I}_q$	$\Sigma_{ij} = 0.2^{\lvert i-j \rvert}$
Non-sparse $r = 1$	$\begin{bmatrix} \mathbf{1}_{3\times 1} \\ \mathbf{0.1}_{8\times 1} \end{bmatrix}$	$a \cdot \begin{bmatrix} \mathbf{1}_{6\times 1} \\ \mathbf{0.1}_{5\times 1} \end{bmatrix}$	$\boldsymbol{\Gamma}_{1,ij} = \begin{cases} \gamma & \text{if } j = i \\ \gamma \cdot 10^{-4} & \text{if } j \neq i \end{cases}$	$\gamma \boldsymbol{I}_q$
	with $a = -0.2, -0.4, \ldots, -0.8$ and $\gamma = 0.4$			

Table 2
Average angle between the estimated and true cointegration spaces.

Method	a							
	-0.2	-0.4	-0.6	-0.8	-0.2	-0.4	-0.6	-0.8
	Low-dimensional				High-dimensional			
	Sparse $q = 4$, $T = 500$, $r = 1$				Sparse $q = 11$, $T = 50$, $r = 1$			
ML	0.032	0.016	0.011	0.008	1.044	0.796	0.559	0.409
PML	**0.020**	**0.010**	**0.007**	**0.005**	**0.588**	**0.226**	**0.160**	**0.138**
	Sparse $q = 4$, $T = 500$, $r = 2$				Sparse $q = 11$, $T = 50$, $r = 4$			
ML	0.007	0.004	0.003	0.002	0.167	0.088	0.058	0.043
PML	**0.006**	**0.003**	**0.002**	**0.001**	**0.138**	**0.065**	**0.041**	**0.029**
	Non-sparse $q = 4$, $T = 500$, $r = 1$				Non-sparse $q = 11$, $T = 50$, $r = 1$			
ML	**0.032**	**0.016**	**0.011**	**0.008**	1.045	0.775	0.542	0.384
PML	0.037	0.019	0.013	0.009	**0.646**	**0.289**	**0.220**	**0.248**

Notes: The results are reported for different values of the adjustment coefficient a and dimension q of the VECM. Differences between the PML and ML estimators that are significant at the 5% level are shown in bold.

the estimated and true cointegration spaces for different values of the adjustment coefficients a. We use a two-sided paired t-test to test equality of the average angle of the PML and ML estimators.

In the sparse low-dimensional settings, the sparse estimator performs the best, providing estimates that are significantly more precise than those of Johansen's estimator for almost all values of the adjustment coefficients. In the non-sparse low-dimensional setting, Johansen's ML estimator performs best, as expected. Using the PML procedure does not lead to a lower estimation precision here.

The advantage of the PML estimator becomes much greater in the high-dimensional designs. The length of the time series is short compared to the number of time series, such that the estimation imprecision of Johansen's ML estimator becomes large. Indeed, the PML estimator outperforms Johansen's ML estimator significantly in all settings, including the non-sparse setting. The differences are large. Since the PML estimator performs regularization, its good performance is retained in the non-sparse high-dimensional setting.

5.3. Forecast accuracy

We evaluate the out-of-sample forecast accuracy using a rolling window of size S. Let h be the forecast horizon. At each time point $t = S, \ldots, T - h$, we use either the PML or Johansen's ML estimator to estimate the VECM

$$\widehat{\Delta \boldsymbol{y}}_{t+h} = \sum_{i=1}^{p-1} \widehat{\boldsymbol{\Gamma}}_i \Delta \boldsymbol{y}_{t+1-i} + \widehat{\boldsymbol{\Pi}} \boldsymbol{y}_t, \tag{9}$$

for different forecast horizons $h \in \{1, 3, 6, 12\}$, obtaining h-step-ahead multivariate forecast errors $\widehat{\boldsymbol{e}}_{t+h} = \Delta \boldsymbol{Y}_{t+h} - \widehat{\Delta \boldsymbol{y}}_{t+h}$. In each simulation run, the overall multivariate forecast performance is then measured using the multivariate mean absolute forecast error (e.g., Carriero et al., 2011):

$$\text{MMAFE} = \frac{1}{T-h-S+1} \sum_{t=S}^{T-h} \frac{1}{q} \sum_{i=1}^{q} \frac{\lvert \Delta y_{t+h}^{(i)} - \widehat{\Delta y}_{t+h}^{(i)} \rvert}{\widehat{\sigma}_{(i)}}, \tag{10}$$

where $\widehat{\sigma}_{(i)}$ is the standard deviation of the ith time series in differences. The MMAFE depends on the forecast horizon h.

Table 3
Multivariate mean absolute forecast errors using the PML and ML estimators: low-dimensional designs.

Setting	$h = 1$		$h = 3$		$h = 6$		$h = 12$	
Window size S	PML	ML	PML	ML	PML	ML	PML	ML
Sparse $q = 4, T = 500, r = 1$								
$S = 48$	**0.85**	0.88	**0.84**	0.88	**0.84**	0.89	**0.84**	0.89
$S = 96$	**0.84**	0.85	**0.83**	0.84	**0.83**	0.85	**0.83**	0.85
$S = 144$	**0.83**	0.84	**0.82**	0.83	**0.82**	0.84	**0.82**	0.84
Sparse $q = 4, T = 500, r = 2$								
$S = 48$	**0.88**	0.97	**0.88**	0.95	**0.88**	0.96	**0.88**	0.96
$S = 96$	**0.87**	0.93	**0.87**	0.91	**0.87**	0.92	**0.87**	0.92
$S = 144$	**0.87**	0.91	**0.87**	0.90	**0.87**	0.90	**0.87**	0.91
Non-sparse $q = 4, T = 500, r = 1$								
$S = 48$	**0.86**	0.89	**0.85**	0.88	**0.85**	0.90	**0.85**	0.89
$S = 96$	**0.85**	0.86	**0.84**	0.85	**0.83**	0.85	**0.84**	0.86
$S = 144$	**0.84**	0.85	**0.83**	0.84	**0.83**	0.84	**0.83**	0.84

Note: the lowest values for each window size S (rows) and forecast horizon h (columns) combination are indicated in bold.

Table 4
Multivariate mean absolute forecast errors using the PML and ML estimators: high-dimensional designs.

Setting	$h = 1$		$h = 3$		$h = 6$		$h = 12$	
	PML	ML	PML	ML	PML	ML	PML	ML
Sparse $q = 11, T = 50, r = 1$	**0.87**	0.91	**0.88**	1.08	**0.87**	1.07	**0.87**	1.06
Sparse $q = 11, T = 50, r = 4$	**0.98**	1.16	**0.99**	1.28	**0.98**	1.26	**0.97**	1.27
Non-sparse $q = 11, T = 50, r = 1$	**0.92**	0.93	**0.93**	1.09	**0.92**	1.08	**0.90**	1.06

Note: the lowest values for each forecast horizon h are indicated in bold.

For the low-dimensional designs, we consider various different window sizes $S \in \{48, 96, 144\}$. The window size S is the number of time points that are available for estimation. We expect the gain in forecast performance of the PML estimator relative to the ML estimator to be larger for small values of S. For the high-dimensional designs, we only consider a window size of $S = 36$ to have sufficient time points available for the estimation of the models.

5.3.1. Results

Simulation results for out-of-sample forecast accuracies in the low-dimensional designs are given in Table 3. For the sake of brevity, we only report the results for $a = -0.4$. The MMAFE is computed for four different forecast horizons (columns) and three rolling window sizes (rows). The PML estimator always obtains lower MMAFE values than the ML estimator. A two-sided paired t-test confirms that these improvements in forecast performance are significant (all p-values <0.01). The forecast accuracy of the PML estimator is also better than that of the ML estimator in the non-sparse low-dimensional setting, though the differences between the two are small, especially for $S = 144$. Regardless of the degree of sparsity of the cointegrating vector (i.e., the number of zero components in the cointegrating vector), the largest gain in forecast accuracy of the PML relative to the ML estimator is obtained when the rolling window size is the lowest ($S = 48$), and this is true for all forecast horizons. Furthermore, the forecast performance of the PML estimator is stable for the different rolling window sizes, while that of the ML estimator varies considerably with the rolling window size.

The simulation results for the forecast accuracy in the high-dimensional designs (for $a = -0.4$) are given in Table 4. The forecast accuracy of the PML estimator is significantly better than that of the ML estimator for all forecast horizons (all p-values <0.01). Overall, the improvements in forecast accuracy are larger for these high-dimensional designs than for the low-dimensional designs in Table 3. The largest forecast accuracy gains of the PML estimator relative to the ML estimator are obtained for the longer forecast horizons.

5.4. Rank determination

We now evaluate the performance of the rank selection criterion (RSC) in selecting the true cointegration rank, and compare it with the trace statistic of Johansen (1988), the Bartlett-corrected trace statistic of Johansen (2002) and the bootstrap procedure of Cavaliere et al. (2012), where the latter two were proposed with the aim of improving on the small-sample performance of Johansen's trace statistic.[4] For each method, we record the relative frequencies of the selected cointegration ranks over all simulation runs.

5.4.1. Results

Table 5 reports the results of the cointegration rank estimation for the low-dimensional designs (for $a = -0.4$). In the first sparse setting, the rank selection criterion performs competitively, with a rank recovery percentage of around 89%. Johansen's method aims to control the size, which results in a rank recovery percentage of around 95% when working with a 5% significance level. Similar results are obtained for the non-sparse low-dimensional setting, and are therefore omitted. In the second sparse setting, RSC selects the cointegration rank correctly in almost all simulation runs.

Table 6 reports the results on the cointegration rank estimation for the high-dimensional designs. RSC performs

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

Table 5
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): low-dimensional designs.

Method	$\hat{r}$									
	0	1	2	3	4	0	1	2	3	4
	Sparse $q = 4, T = 500, r = 1$					Sparse $q = 4, T = 500, r = 2$				
Johansen	0.0	95.4	4.2	0.4	0.0	0.0	0.0	96.2	3.8	0.0
Bartlett	0.0	96.0	3.6	0.4	0.0	0.0	0.0	95.4	4.4	0.2
Bootstrap	0.0	96.8	2.8	0.4	0.0	0.0	0.0	97.2	2.8	0.0
RSC	0.0	89.4	10.6	0.0	0.0	0.0	0.0	98.8	1.2	0.0

Table 6
Frequency of the estimated cointegration rank $\hat{r} = 0, \ldots, q$: high-dimensional designs.

Method	$\hat{r}$											
	0	1	2	3	4	5	6	7	8	9	10	11
	Sparse $q = 11, T = 50, r = 1$											
Johansen	0.0	0.0	0.0	1.0	9.0	15.2	52.0	14.0	7.0	1.6	0.2	0.0
Bartlett	0.0	11.2	31.8	20.2	14.0	6.6	6.2	4.4	3.8	1.6	0.2	0.0
Bootstrap	98.8	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RSC	0.0	57.4	40.4	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Sparse $q = 11, T = 50, r = 4$											
Johansen	0.0	0.0	0.0	3.2	24.6	23.4	41.4	6.4	0.8	0.2	0.0	0.0
Bartlett	0.0	7.6	18.4	23.6	19.0	11.8	10.0	5.4	3.2	0.8	0.2	0.0
Bootstrap	99.0	0.8	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RSC	0.0	0.0	9.0	60.6	28.8	1.6	0.0	0.0	0.0	0.0	0.0	0.0

much better than its alternatives in all settings. In the first setting, RSC estimates the cointegration rank correctly in 57.4% of the simulation runs, the Bartlett-corrected trace statistic in 11.2%, the bootstrap in 1.2% and Johansen's trace statistic in 0%. 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 1%, for instance. Similar results are obtained for the non-sparse setting.

When the true cointegration rank increases ($r = 4$ in the second setting), the performance of the rank selection criterion becomes sensitive to the strength of the cointegration signal: its rank recovery percentage increases from 28.8% for $a = -0.4$ to 73.8% for $a = -0.8$ (unreported). However, RSC still performs the best.

In contrast to Johansen's trace statistic, the RSC is not meant to control the size. One must take into account the difficulty of comparing size-targeting methods, such as Johansen's trace statistic, with consistency-targeting methods, such as the RSC, when assessing the results on cointegration rank determination. The RSC also has the tendency to overestimate the cointegration rank rather than underestimating it. Overestimation is less severe, since the PML estimator allows some of the cointegrating vectors, i.e., columns of $\boldsymbol{\beta}$, to be estimated as zero. Then, the actual rank of $\hat{\boldsymbol{\beta}}$ will be lower than that estimated by the RSC.

6. Forecasting

We evaluate the forecast performance of the sparse cointegration method on two data sets. In the first data set, we have interest rates of different maturities. Financial theory implies that these interest rates of different maturities will be cointegrated. We consider a VECM and compare the forecast performances of the sparse cointegration method and the traditional method. For the second data set, we forecast a large number of industry-specific consumption time series. We investigate the question of whether the forecast accuracy can be improved by using the sparse cointegration method rather than alternative methods.

We evaluate the forecast accuracy by performing rolling window forecasting, as described in Section 5.3. We use the rank selection criterion from Section 4 to estimate the cointegration rank, and the BIC to select the order p of the VECM. In addition to the multivariate mean absolute forecast error, we also provide results for predicting the individual time series $\Delta y_t^{(i)}$, $i = 1, \ldots, q$, by computing the mean absolute forecast error

$$\text{MAFE} = \frac{1}{T-h-S+1} \sum_{t=S}^{T-h} \frac{|\Delta y_{t+h}^{(i)} - \widehat{\Delta y}_{t+h}^{(i)}|}{\widehat{\sigma}_{(i)}}. \tag{11}$$

We compare the forecast performances of the different methods using the Diebold–Mariano test (DM-test, see Diebold & Mariano, 1995).

6.1. Interest rate growth forecasting

In finance, the expectations hypothesis of interest rates (e.g., Engsted & Tanggaard, 1994; Giese, 2008) implies that the interest rates of different maturities will be cointegrated. We collect monthly data on $q = 5$ US treasury bills with different times to maturity, ranging from July 1969 to June 2015, giving $T = 552$ (source: Datastream, Federal Reserve, US). A time plot of the interest rates is provided in Fig. 1. All of the interest rates move very closely together, meaning that we would expect them to be cointegrated. A stationarity test of all individual interest rates using the Augmented Dickey–Fuller test confirms that the time series are integrated of order 1. We

Table 7
Multivariate mean absolute forecast error using the PML and ML estimators.

Window size	$h = 1$		$h = 3$		$h = 6$		$h = 12$	
	PML	ML	PML	ML	PML	ML	PML	ML
$S = 48$	**0.70**	1.13***	**0.70**	0.86***	**0.74**	0.98***	**0.70**	0.86***
$S = 96$	**0.68**	0.84**	**0.68**	0.74***	**0.69**	0.77***	**0.70**	0.75**
$S = 144$	**0.63**	0.71**	**0.61**	0.66**	**0.59**	0.65***	**0.58**	0.65**

Note: the lowest values for each window size S (rows) and forecast horizon h (columns) combination are indicated in bold. For the DM-test of equal MMAFEs of the two methods:
* Indicate significance at the 10% level.
** Indicate significance at the 5% level.
*** Indicate significance at the 1% level.

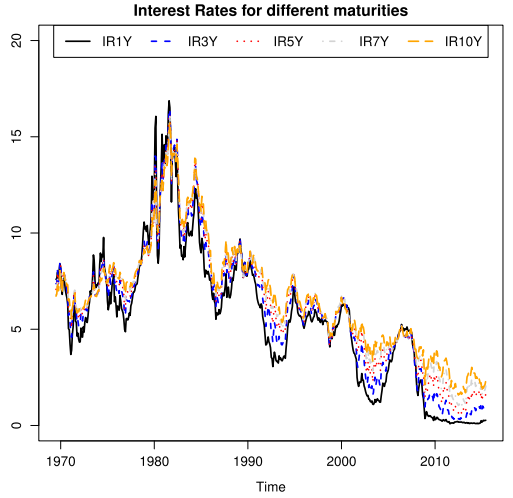

Fig. 1. Time plot (July 1969–June 2015) of the interest rates for the different maturities: one year (black solid line), three years (blue short-dashed line), five years (red dotted line), seven years (gray dot-dashed line), ten years (orange long-dashed line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

take the cointegration relationships implied by financial theory into account by estimating a VECM with q interest rates.

We investigate the behavior of the penalized maximum likelihood estimator compared to that of the Johansen maximum likelihood estimator when the length of the time series varies relative to the fixed dimension $q = 5$. For this purpose, we consider three different window sizes: $S \in \{48, 96, 144\}$.

The multivariate mean absolute forecast error is computed for four different forecast horizons (columns) and three different rolling window sizes (rows), see Table 7. The PML estimator beats Johansen's estimator in all settings, and a DM-test confirms that this improvement in forecast performance is significant overall. The MMAFE of the PML estimator remains relatively stable as the window size varies, whereas that of the ML estimator becomes much worse as the window size shrinks. For a window size of $S = 48$, the MMAFE of the PML estimator is 25% lower than that of Johansen's estimator, on average. As the window size increases, the PML estimator still performs the best, but the difference between the two becomes somewhat smaller.

The mean absolute forecast errors for the five individual interest rate time series are reported in Table 8. The PML estimator delivers the most accurate forecasts for all interest rates, forecast horizons and window sizes considered. The largest forecast accuracy gains occur for the smallest window size S.

In summary, when the time series length is short compared to the number of time series to be predicted, important forecast accuracy gains can be obtained by using the sparse estimator instead of the Johansen's estimator. However, sparsity leads to improvements in forecast accuracy for real data even for long time series, since the sparse estimator delivers a more parsimonious model.

6.2. Consumption growth forecasting

Our objective is to predict a large number of industry-specific consumption time series. We collect monthly data on $q = 31$ US consumption time series, between January 1999 and April 2015, thus giving $T = 196$ (see Table 11, Appendix B for a data description). Personal consumption accounts for around 70% of GDP in the US, and is monitored closely by public policy makers and marketing managers (Fornell, Rust, & Dekimpe, 2010). In contrast to total consumption, industry-specific consumption time series have often been discarded previously in the forecasting literature, as they are typically highly collinear, which might create estimation problems (Carriero et al., 2011). We exploit the co-movement among these time series by forecasting the total and industry-specific consumption growth in a cointegration framework using the PML estimator from Section 3. Time plots of all log-transformed consumption time series are provided in Fig. 2 of Appendix B. A stationarity test of all individual log-transformed time series using the Augmented Dickey–Fuller test confirms that they are integrated of order one, and we forecast consumption growth using a VECM.

We conduct a rolling window forecast exercise using a window of 12 years of data ($S = 144$), and compare the performances of eight estimators. The first three estimators are estimators for the (log-transformed) consumption time series that account for cointegration, while the remainder are estimators for the consumption growth time series that do not account for cointegration. The estimators are (1) PML estimation of the VECM (see Section 3), (2) ML estimation of the VECM, (3) the factor model of Barigozzi, Lippi, and Luciani (2016) for non-stationary time series, (4) PML estimation of the VAR, (5) ML estimation of the VAR, (6) the factor model of Stock and Watson (2002) for

Table 8
Mean absolute forecast errors for the $q = 5$ individual interest rate time series using the PML and ML estimators.

Window size	Interest rate	$h = 1$		$h = 3$		$h = 6$		$h = 12$	
		PML	ML	PML	ML	PML	ML	PML	ML
$S = 48$	1Y	**0.60**	0.73***	**0.61**	0.74***	**0.62**	0.76***	**0.60**	0.66**
	3Y	**0.66**	0.99***	**0.67**	0.86***	**0.68**	0.92***	**0.67**	0.87***
	5Y	**0.70**	1.25***	**0.73**	0.92***	**0.77**	1.01***	**0.73**	0.88***
	7Y	**0.73**	1.46***	**0.72**	0.89***	**0.75**	1.04***	**0.74**	0.92***
	10Y	**0.81**	1.19**	**0.79**	0.89*	**0.88**	1.15*	**0.78**	0.97***
$S = 96$	1Y	**0.54**	0.57	**0.55**	0.59**	**0.56**	0.59	**0.55**	0.57
	3Y	**0.66**	0.80**	**0.66**	0.72***	**0.66**	0.72***	**0.68**	0.73**
	5Y	**0.70**	0.92**	**0.70**	0.78**	**0.72**	0.83**	**0.70**	0.79***
	7Y	**0.73**	1.01**	**0.73**	0.78***	**0.74**	0.84***	**0.74**	0.79***
	10Y	**0.78**	0.91*	**0.75**	0.84***	**0.78**	0.88**	**0.80**	0.88
$S = 144$	1Y	**0.44**	0.48***	**0.43**	0.45	**0.41**	0.45**	**0.40**	0.43***
	3Y	**0.59**	0.68***	**0.59**	0.65**	**0.57**	0.62**	**0.56**	0.60***
	5Y	**0.64**	0.79*	**0.64**	0.71*	**0.63**	0.69***	**0.60**	0.66**
	7Y	**0.69**	0.82*	**0.68**	0.73*	**0.66**	0.73***	**0.64**	0.76*
	10Y	**0.78**	0.78	**0.72**	0.74	**0.69**	0.77**	**0.69**	0.80

Note: the lowest values for each interest rate and window size-forecast horizon combination are indicated in bold. For the DM-test of equal MAFEs of the two methods:.
* Indicate significance at the 10% level.
** Indicate significance at the 5% level.
*** Indicate significance at the 1% level.

Table 9
Multivariate mean absolute forecast errors (MMAFE) for the different methods (columns) and forecast horizons h (rows).

Forecast horizon	Cointegration			No cointegration				
	PML	ML	Factor model	PML	ML	Factor model	Bayesian	Bayesian reduced rank
$h = 1$	0.79	0.74	**0.66**	0.94	5.40***	0.72	0.69	0.69
$h = 3$	**0.62**	0.78***	0.66***	0.67***	4.81***	0.75***	0.71***	0.71***
$h = 6$	**0.63**	0.82***	0.67***	0.67***	4.84***	0.77***	0.74***	0.74***
$h = 12$	**0.61**	0.72***	0.65***	0.66***	5.22***	0.72***	0.72***	0.72***

For the DM-test of equal MMAFEs of a given method and the PML method for cointegration:
* Indicate significance at the 10% level.
** Indicate significance at the 5% level.
*** Indicate significance at the 1% level.

stationary time series, (7) Bayesian estimation of the VAR with the Normal-Inverse Wishart prior introduced by Banbura et al. (2010), and (8) Bayesian reduced rank regression (Carriero et al., 2011), which combines the benefits of rank reduction and Bayesian shrinkage.[5] Note that the forecast performances being always evaluated in terms of MMAFEs or MAFEs are for the time series in *differences*. As a result, the forecast errors of the different estimators are comparable. We have also included an intercept in the VECM of Eq. (1), since some of the consumption time series exhibit drifts.

The multivariate mean forecast errors are reported in Table 9. The PML estimator of the VECM obtains the lowest value for all forecast horizons except for $h = 1$.[6] A DM-test confirms that the differences in forecast performance are significant. Taking the long-run cointegration relationships into account pays off, especially for the longer forecast horizons. Taking cointegration into account (PML, ML, factor model) yields significantly better forecasts than not accounting for cointegration in almost all cases. Of the methods that account for cointegration, the PML estimator performs best, thus confirming that sparse estimation improves the forecast performance. The PML estimator of the VECM also performs significantly better than the Bayesian estimators.

Individual mean absolute forecast errors for the separate time series are also computed. For the sake of brevity, Table 10 only reports them for the total consumption time series. The results for the MAFE are similar to those of the MMAFE. The PML and ML estimators and the factor model that account for cointegration attain (significantly) better MAFEs than the corresponding methods that do not account for it. The proposed PML estimator of the VECM obtains the best value of the MAFE for all forecast horizons except for $h = 1$.

In summary, the sparse cointegration method is a valuable addition to the forecaster's toolbox for high-dimensional time series. It exploits the co-movements

[5] For estimators (3), (7) and (8), the rank and the number of factors k are determined by calculating the maximum eigenvalue ratio criterion $\hat{k}_j = \hat{\lambda}_j/\hat{\lambda}_{j+1}$, for $j = 1, \ldots, q-1$, from the eigenvalues $\hat{\lambda}_j, \ldots, \hat{\lambda}_q$ and selecting $k = \text{argmax}_j \hat{k}_j$.

[6] Although the factor model for cointegration obtains the best MMAFE for $h = 1$, its forecast performance is not significantly different from that of the PML method for cointegration.

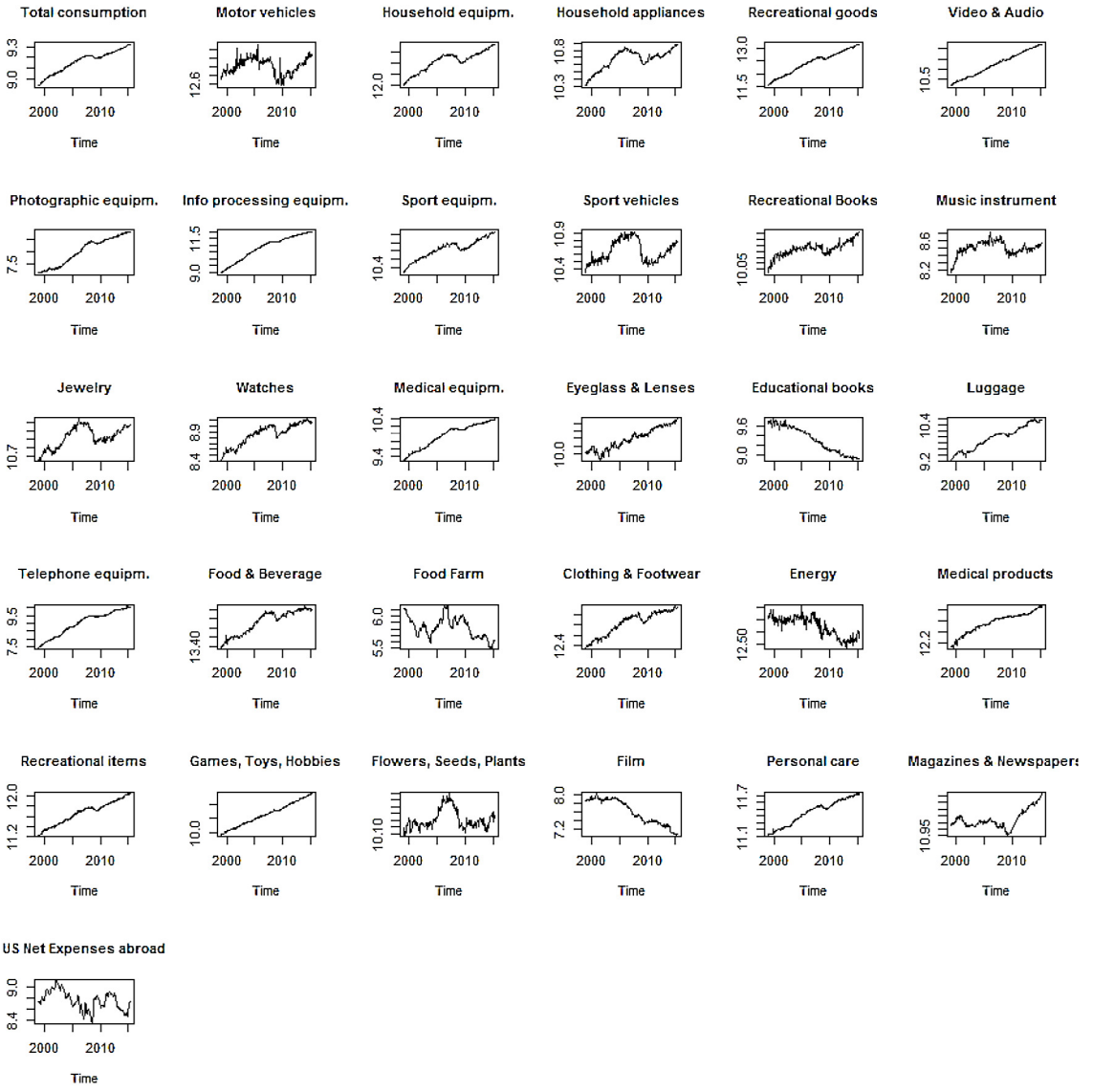

Fig. 2. Time plot (January 1999–April 2015) of the total consumption time series, the 18 durable consumption time series, and the 12 nondurable consumption time series, all in logs.

Table 10
Mean absolute forecast errors (MAFE) for the total consumption time series, for different methods (columns) and forecast horizons h (rows).

Forecast horizon	Cointegration			No cointegration				
	PML	ML	Factor model	PML	ML	Factor model	Bayesian	Bayesian reduced rank
$h = 1$	3.82	0.61	**0.59**	6.14	47.28***	0.59	0.65	0.66
$h = 3$	**0.46**	0.66***	0.59***	0.59***	44.44***	0.58*	0.57**	0.57**
$h = 6$	**0.48**	0.81***	0.60**	0.60**	43.96***	0.76***	0.71***	0.71***
$h = 12$	**0.46**	0.62***	0.61***	0.61***	57.61***	0.64**	0.80***	0.79***

See the notes to Table 9.

among large numbers of time series by estimating the cointegration relationships sparsely.

7. Conclusion

This paper has discussed a sparse cointegration method. Our simulation study shows that the sparse method outperforms Johansen's ML method significantly if the true cointegrating vectors are sparse or if the time series length is short compared to the number of time series. The degree of sparsity that is needed in order for the sparse estimator to outperform the ML estimator depends on the time series length relative to the number of time series. The higher the

degree of sparsity, the more quickly the sparse estimator will outperform the ML estimator.

Sparse cointegration methods are useful for several reasons. In high-dimensional settings with cointegrated time series, estimating the cointegrating vectors sparsely might improve the estimation accuracy and/or forecast performance. We show that the sparse cointegration method achieves important gains in forecast accuracy compared to the traditional maximum likelihood estimator if the time series length is short compared to the number of time series (cfr. interest rate forecasting). When forecasting highly collinear time series (cfr. consumption forecasting), important gains can be obtained by accounting for cointegration and by estimating the cointegration relations sparsely.

The sparse cointegration method might suffer from the following points. First, we impose the normalization condition on $\boldsymbol{\alpha}$ rather than on $\boldsymbol{\beta}$. As such, the weighted Procrustes problem might be affected by multicollinearity issues. In addition, we also impose sparsity on $\boldsymbol{\beta}$, which is not defined uniquely. This might pose difficulties for model interpretation. However, the consequences of these issues for the forecast performance of the proposed method are less severe.

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). While Johansen's trace statistic cannot be computed once the total number of lagged time series $(p-1)\cdot q$ exceeds the time series length T, the rank selection criterion, as presented in Section 4, requires the number of time series q to be smaller than the time series length T. Further research is needed to determine how to improve its implementation for truly high-dimensional settings where $q > T$. The eigenvalue-ratio-based rank estimator of Lam and Yao (2012) might be an alternative to the RSC for such settings.

There are several questions that we have not addressed but have left for future research. For instance, the models analyzed in this paper generally exclude deterministic terms (Nielsen & Rahbek, 2000). We also excluded structural breaks, although allowing for structural breaks can be useful when analyzing economic data (Johansen, Mosconi, & Nielsen, 2000). A natural extension of this study would be to implement structural analysis; for instance, impulse–response functions can be estimated using the PML estimator. Confidence bounds around the impulse–response functions can then be obtained using a bootstrap procedure.

Acknowledgment

The authors gratefully acknowledge financial support from the FWO (Research Foundation Flanders, contract number 11N9913N).

Appendix A. Time series cross-validation

We select the tuning parameters following a time series cross-validation approach (Hyndman, 2014). Denote the response by $\boldsymbol{z}_t$. When solving for $\boldsymbol{\Gamma}$, $\boldsymbol{z}_t = \Delta\boldsymbol{y}_t - \boldsymbol{\Pi}\boldsymbol{y}_{t-1}$. When solving for $\boldsymbol{\Pi}$, $\boldsymbol{z}_t = \Delta\boldsymbol{y}_t - \sum_{i=1}^{p-1}\boldsymbol{\Gamma}_i\Delta\boldsymbol{y}_{t-i}$.

1. For $t = S, \ldots, T-1$ (with $S = \lfloor 0.8T \rfloor$), repeat:
   (a) For a grid of tuning parameters, fit the model to the data $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_t$.
   (b) Compute the one-step-ahead forecast error $\widehat{\boldsymbol{e}}_{t+1} = \boldsymbol{z}_{t+1} - \hat{\boldsymbol{z}}_{t+1}$.
2. Select the value of the tuning parameter that minimizes the mean squared forecast error

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

where $\hat{e}_t^{(i)}$ is the ith component of the multivariate time series at time t and $\hat{\sigma}_{(i)}$ is the standard deviation of the time series $z_t^{(i)}$.

Appendix B. Consumption time series

See Table 11.

Table 11
Consumption expenditures.
Source: Datastream, Bureau of Economic Analysis.

Total consumption
Durable consumption: Motor vehicles and parts
Durable consumption: Furnishings and durable household equipment
Durable consumption: Household appliances
Durable consumption: Recreational goods and vehicles
Durable consumption: Video and audio equipment
Durable consumption: Photographic equipment
Durable consumption: Information processing equipment
Durable consumption: Sporting equipment, supplies, guns and ammunition
Durable consumption: Sports and recreational vehicles
Durable consumption: Recreational books
Durable consumption: Musical instruments
Durable consumption: Jewelry
Durable consumption: Watches
Durable consumption: Therapeutic medical equipment
Durable consumption: Corrective eyeglasses and contact lenses
Durable consumption: Educational books
Durable consumption: Luggage
Durable consumption: Telephone equipment
Nondurable consumption: Food and beverages
Nondurable consumption: Food produced and consumed on farms
Nondurable consumption: Clothing and footwear
Nondurable consumption: Gasoline and other energy goods
Nondurable consumption: Pharmaceutical and other medical products
Nondurable consumption: Recreational items
Nondurable consumption: Games, toys and hobbies
Nondurable consumption: Flowers, seeds and potted plants
Nondurable consumption: Film and photographic supplies
Nondurable consumption: Personal care products
Nondurable consumption: Magazines and newspapers
Nondurable consumption: Net expenditure abroad by US residents

Appendix C. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.ijforecast.2016.04.005.

References

Anderson, T. (1958). *An introduction to multivariate statistical analysis*. New York: John Wiley & Sons, Inc..

Banbura, M., Giannone, D., & Reichlin, L. (2010). Large Bayesian vector autoregressions. *Journal of Applied Econometrics*, *25*(1), 71–92.

Barigozzi, M., Lippi, M., & Luciani, M. (2016). Non-stationary dynamic factor models for large datasets. arXiv:1602.0239v1.

Bernardini, E., & Cubadda, G. (2015). Macroeconomic forecasting and structural analysis through regularized reduced-rank regression. *International Journal of Forecasting*, *31*(3), 682–691.

Bunea, F., She, Y., & Wegkamp, M. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, *39*(2), 1282–1309.

Carriero, A., Kapetanios, G., & Marcellino, M. (2011). Forecasting large datasets with Bayesian reduced rank multivariate models. *Journal of Applied Econometrics*, *26*(5), 735–761.

Cavaliere, G., Rahbek, A., & Taylor, A. R. (2012). Bootstrap determination of the co-integration rank in vector autoregressive models. *Econometrica*, *80*(4), 1721–1740.

Chen, L., & Huang, J. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, *107*(500), 1533–1545.

Diebold, F., & Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*(3), 253–263.

Engle, R., & Granger, C. (1987). Cointegration and error correction: Representation, estimation, and testing. *Econometrica*, *55*(2), 251–276.

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

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

Fornell, C., Rust, R., & Dekimpe, M. (2010). The effect of customer satisfaction on consumer spending growth. *Journal of Marketing Research*, *47*(1), 28–35.

Friedman, J. (2012). Fast sparse regression and classification. *International Journal of Forecasting*, *28*(3), 722–738.

Friedman, J., Hastie, T., & 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.

Hamilton, J. (1991). *Time series analysis*. Princeton University Press.

Hyndman, R. (2014). forecast: Forecasting functions for time series and linear models. R package version 5.2. URL: http://cran.r-project.org/package=forecast.

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

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

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

Johansen, S. (2002). A small sample correction of the test for cointegration rank in the vector autoregressive model. *Econometrica*, *70*(5), 1929–1961.

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

Korobilis, D. (2013). VAR forecasting using Bayesian variable selection. *Journal of Applied Econometrics*, *28*(2), 204–230.

Lam, C., & Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics*, *40*(2), 694–726.

Liao, Z., & Phillips, P. (2015). Automated estimation of vector error correction models. *Econometric Theory*, *31*, 581–646.

Lissitz, R., Schonemann, P., & 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.

Lütkepohl, 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), 1–23.

Nielsen, B., & Rahbek, A. (2000). Similarity issues in cointegration models. *Oxford Bulletin of Economics and Statistics*, *62*(1), 5–22.

Phillips, P., & Ouliaris, S. (1990). Asymptotic properties of residual based tests for cointegration. *Econometrica*, *58*(1), 165–193.

Rothman, A., Levina, E., & Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, *19*(4), 947–962.

Stock, J., & Watson, M. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics*, *20*(2), 147–162.

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

Zhou, X., Nakajima, J., & West, M. (2014). Bayesian forecasting and portfolio decisions using dynamic sparse factor models. *International Journal of Forecasting*, *30*(4), 963–980.

Ines Wilms joined the Research Centre for Operations Research and Business Statistics at KU Leuven in September 2012 as a doctoral researcher. Her research is supervised by Christophe Croux and she obtained an FWO Aspirant research grant from the Research Foundation Flanders. She is interested in time series analysis, forecasting and multivariate statistics. Currently, the main focus of her research is on the development of sparse and robust methods for estimating relationships between large numbers of time series.

Christophe Croux is Full Professor at KU Leuven, Belgium. He is interested in time series analysis, forecasting, robust methods, computational statistics, exploratory data analysis and mathematical statistics. Currently, he studies variable selection problems and the robust estimation of high-dimensional problems. The application and implementation of the proposed methodology form an essential part of his research projects.