International Journal of Forecasting 34 (2018) 408–430

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Macroeconomic forecasting using penalized regression methods

Stephan Smeekes, Etienne Wijler *

Maastricht University, Department of Quantitative Economics, The Netherlands

ARTICLE INFO

Keywords:
Forecasting
Lasso
Factor models
High-dimensional data
Cointegration

ABSTRACT

We study the suitability of applying lasso-type penalized regression techniques to macroeconomic forecasting with high-dimensional datasets. We consider the performances of lasso-type methods when the true DGP is a factor model, contradicting the sparsity assumptionthat underlies penalized regression methods. We also investigate how the methods handle unit roots and cointegration in the data. In an extensive simulation study we find that penalized regression methods are more robust to mis-specification than factor models, even if the underlying DGP possesses a factor structure. Furthermore, the penalized regression methods can be demonstrated to deliver forecast improvements over traditional approaches when applied to non-stationary data that contain cointegrated variables, despite a deterioration in their selective capabilities. Finally, we also consider an empirical applicationto a large macroeconomic U.S. dataset and demonstrate the competitive performance of penalized regression methods.

1. Introduction

In this paper we provide a thorough analysis of the forecasting capabilities of penalized regression in macroeconomic conditions. We study the performance of these methods in a simulation study when the true DGP is a factor model and when the data contain stochastic trends and may be cointegrated. We also provide a systematic comparison with factor models, the mainstream method used in macroeconomic forecasting, using both Monte Carlo simulations and an empirical application to macroeconomic data.

Despite the vast size of the forecasting literature, comprehensive comparisons between factor models and penalized regression remain scarce. Traditionally, the majority of the forecasting literature seems to have implicitly assumed the prevalence of a latent factor structure in economic datasets and therefore has mainly considered the performance of methods based on factor estimation. While very popular in statistics, only recently ℓ_1-penalized regression techniques, such as the lasso from Tibshirani (1996), are being explored as a viable alternative in macroeconometrics. Applications in forecasting in particular show that the use of penalized regression, potentially in combination with traditional techniques such as principal components (PC), delivers promising performance (e.g Kim & Swanson, 2014), though it is not yet really understood why. By providing a comprehensive study of penalized regression in "adverse" macroeconomic conditions, we complement the existing literature with a fresh perspective on these methods and a direct link to factor models.

Specifically, we address the apparent contradiction between the premise of forecasting with shrinkage estimators to identify a small subset of variables responsible for the variation in the dependent variable and the assumption that the variation in the dependent variable is best explained through aggregates of all available time series. The good empirical performance of penalized regression methods despite this contradiction gives rise to a number of practically relevant questions; (1) Is the common factor assumption really valid in practice? (2) Are the results due

* Correspondence to: Department of Quantitative Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands.
E-mail address: E.Wijler@maastrichtuniversity.nl (E. Wijler).

https://doi.org/10.1016/j.ijforecast.2018.01.001
0169-2070/

to sample-dependent data idiosyncrasies? (3) Are other mechanisms at play such as an inherent robustness of shrinkage estimators to alternative DGP specifications?

We aim to shed light on these previously unexplored questions by conducting a detailed simulation study in which we compare the performance of a selection of the most popular and well understood variants of ℓ_1-shrinkage estimators and factor extraction methods. The novelty in these simulations comes from the wide range of DGPs considered, chosen such that no method is consistently favoured over another based on a priori expectations and to closely resemble the types of data that occur in empirical applications. The former goal is maintained through varying both the presence of common factors in the data as well as the degree of sparsity in the parameter space, while the latter goal is maintained through introducing levels of non-sphericity frequently encountered in empirical work.[1] In addition, we explore the potential of penalized regression in the non-stationary setting by generating a number of time series containing unit roots, some of which are cointegrated, and employ penalized regression directly on these series without any form of preprocessing. We complement the simulations with a comparison of the pseudo out-of-sample forecasting performance on a recently updated U.S. macroeconomic dataset available through the FredMD database (McCracken & Ng, 2015).

The results show that penalized regression performs remarkably well when there is at least some degree of sparsity in the parameter space and is relatively robust against alternative DGP specifications. Factor models perform slightly better than penalized regression when the predictors possess an approximate factor structure with low dependence in the errors, but their performance deteriorates substantially when increasing the level of non-sphericity in the idiosyncratic component. Penalized regression naturally does better than factor models on DGPs without factors, but more surprisingly also provides forecast improvements on DGPs containing a factor structure with strongly serially and cross-sectionally correlated idiosyncratic components. In addition, penalized regression shows promising results on cointegrated data, producing substantially lower forecast errors compared to standard OLS despite failing to identify the exact cointegrating vector at relatively high frequencies. Finally, the empirical application highlights that the forecast performance differentials between factor-based methods and shrinkage methods are sensitive to the target variable being forecast.

Our contribution complements the vast existing macroeconomic forecasting literature that is dominated by methods that exploit a latent factor structure, such as static factor models (e.g. Bai & Ng, 2008; Stock & Watson, 2002a,b), dynamic factor models (Doz, Giannone, & Reichlin, 2012; Eickmeier & Ziegler, 2008; Forni, Giovannelli, Lippi, & Soccorsi, 2016; Forni, Hallin, Lippi, & Reichlin, 2005), weighted principal components (Boivin & Ng, 2006), sparse principal components (Kristensen, 2017) or factor augmented vector autoregressions (Bai, Li, & Lu, 2016; Bernanke, Boivin, & Eliasz, 2005; Pesaran, Pick, & Timmerman, 2011). The conjecture that a small set of factors drives the variation in economic time series finds strong support through impressive forecasting performance of factor models on macroeconomic datasets from the U.S. (Stock & Watson, 2002a, 2012), the U.K. (Artis, Banerjee, & Marcellino, 2005) and the Euro area (Marcellino, Stock, & Watson, 2003). Spurred by theoretical developments such as the extension of the adaptive lasso to general time series frameworks by Medeiros and Mendes (2016), ℓ_1-penalized regression has gained more appeal and the body of applied literature taking into account these shrinkage estimators has grown considerably, with recent work covering penalized regression (De Mol, Giannone, & Reichlin, 2008; Gelper & Croux, 2008; Kim & Swanson, 2014; Li & Chen, 2014), reduced-rank vector autoregressions (Bernardini & Cubadda, 2015), Bayesian vector autoregressions (Bańbura, Giannone, & Reichlin, 2010) and penalized vector autoregressions (Barigozzi & Brownlees, 2017; Callot & Kock, 2014; Hsu, Hung, & Chang, 2008; Kascha & Trenkler, 2015). While some include a direct comparison between at least some form of factor models and penalized regression and demonstrate predictive capabilities of ℓ_1-penalized regression that is competitive to traditional factor models, the analysis is typically based on empirical data or simulations that do not provide detailed insights into the sensitivity of each method to its underlying assumptions.

The remainder of this paper is organized as follows. Section 2 describes the notation and reviews the methods considered. In Section 3 we perform the simulation based analysis of the forecasting performance, followed by the empirical application in Section 4. In Section 5 we conclude and suggest a number of interesting avenues for future research.

2. Methods

Suppose a researcher is interested in predicting an economic time series h-steps ahead with information available through time $t = 1, \ldots, T$. The researcher desires to include a pre-determined set of variables such as lags of the dependent variable or variables motivated through economic theory. In addition, she faces a large set of candidate variables that are potentially relevant to the dependent variable. This results in the following generic model:

$$y_{t+h} = \boldsymbol{w}_t'\boldsymbol{\beta}_w + \boldsymbol{x}_t'\boldsymbol{\beta}_x + \epsilon_{t+h} \tag{1}$$

where y_{t+h} is the scalar valued dependent variable to forecast and h is the forecast horizon. $\boldsymbol{w}_t$ is the $(p \times 1)$ predetermined vector of variables which the researcher requires to be in the model, $\boldsymbol{x}_t$ is the $(N \times 1)$ vector containing candidate variables that are potentially related to y_{t+h}, and ϵ_{t+h} is a disturbance term. The forecast of the response at time T is defined as $\hat{y}_{T+h|T} = \boldsymbol{w}_T'\hat{\boldsymbol{\beta}}_w + \boldsymbol{x}_T'\hat{\boldsymbol{\beta}}_x$. Letting $\boldsymbol{y} = (y_{1+h}, \ldots, y_{T+h})'$, $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T)'$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)'$ and $\boldsymbol{\epsilon} = (\epsilon_{1+h}, \ldots, \epsilon_{T+h})$ the model can be rewritten as

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{\beta}_w + \boldsymbol{X}\boldsymbol{\beta}_x + \boldsymbol{\epsilon}. \tag{2}$$

When the number of variables in the candidate set $\boldsymbol{X}$ is large relative to the number of available observations,

[1] Throughout this paper the term non-sphericity refers to the presence of cross-sectional and/or serial correlation in the idiosyncratic component of any data generating process.

modelling the dependent variable as a linear combination of all candidate variables will amount to the estimation of a large number of parameters and is likely to result in a large forecasting variance. For example, assuming the explanatory variables follow a Gaussian distribution, Stock and Watson (2006) show that the OLS forecast is normally distributed with a variance proportional to the number of variables included in the model divided by the total number of available observations. In the more extreme case where the cross-section dimension exceeds the time series dimension inverting the matrix of second moments becomes infeasible and as a result the OLS estimator does not have a (unique) solution. Accordingly, methods that perform regularization are required in order to obtain accurate forecasts and reliable model estimates in the high-dimensional setting.

The methods we consider can broadly be categorized as shrinkage estimators and factor models. Shrinkage estimators aim to reduce the forecast variance by shrinking the parameter estimates in the traditional linear model, possibly up to a point where some parameters are exactly equal to zero and, thus, removing the corresponding variables from the candidate set. Factor models, on the other hand, do not remove variables from the candidate set, but rather aim to reduce the dimensionality of the data by summarizing the data in relatively few factors with the hope of capturing the bulk of the variation in the candidate set. In the following section we formally introduce these methods and describe the mechanisms by which they estimate our generic model (1).

2.1. Shrinkage estimators

The shrinkage estimators employed in this paper estimate the parameters according to the following objective function:

$$(\hat{\boldsymbol{\beta}}_w, \hat{\boldsymbol{\beta}}_x) = \underset{(\boldsymbol{\beta}_w, \boldsymbol{\beta}_x)}{\arg\min} \sum_{t=1}^{T} (y_{t+h} - \boldsymbol{w}_t'\boldsymbol{\beta}_w - \boldsymbol{x}_t'\boldsymbol{\beta}_x)^2 + \lambda \left[\alpha \sum_{j=1}^{N} \frac{|\beta_{x,j}|}{\omega_j} + (1-\alpha) \sum_{j=1}^{N} \frac{|\beta_{x,j}|^2}{\omega_j} \right], \quad (3)$$

with different settings of $(\lambda, \alpha, \omega_j)$ leading to various well-established methods. We consider:

1. Ridge regression (ridge: $\lambda > 0, \alpha = 0, \omega_j = 1\ \forall j$)
2. Lasso (las: $\lambda > 0, \alpha = 1, \omega_j = 1$),
3. Adaptive Lasso (adalas: $\lambda > 0, \alpha = 1, \omega_j = \left|\hat{\beta}_{Init,j}\right|$),
4. Elastic Net (en: $\lambda > 0, 0 < \alpha < 1, \omega_j = 1\ \forall j$), and
5. Adaptive Elastic Net (adaen: $\lambda > 0, 0 < \alpha < 1, \omega_j = \left|\hat{\beta}_{Init,j}\right|$),

where $\hat{\beta}_{Init,j}$ is an initial estimator such as the OLS or ridge coefficient. The methods that use $\alpha \in (0, 1]$ and $\lambda > 0$, from here on referred to as lasso-type estimators, perform subset selection by shrinking coefficient estimates to zero and, hence, are potentially able to improve forecasting performance by reducing the added variance of estimating parameters of irrelevant variables. Additionally, these methods allow for model estimation in situations where the number of potentially relevant variables exceeds the number of observations, i.e. $N > T$. The weights ω_j, $j = 1, \ldots, N$, allow for differential shrinkage on the parameters. Zou (2006) demonstrates that the use of cleverly chosen initial estimators as weights improves the selection performance by penalizing irrelevant variables to a higher degree than relevant variables. Common choices for initial estimators are the absolute values of OLS or ridge coefficients from a preceding estimation. Furthermore, it can be directly observed from (3) that the pre-determined set of relevant variables $\boldsymbol{w}_t$ is free of regularization and is therefore ensured to be included in the final model. Following Friedman, Hastie, and Tibshirani (2010) the solution to (3) can be efficiently obtained using a coordinate descent algorithm.

Whereas the earlier theory for the lasso has been developed in rather restrictive frameworks such as fixed designs (e.g. Knight & Fu, 2000; Zou, 2006), the properties of the lasso and its variants are becoming increasingly well understood in time series settings. One strand of time series related literature focusses on a framework with a fixed number of independent variables. This includes, among others, the work of Wang, Li, and Tsai (2007) who apply the (adaptive) lasso to models with autoregressive errors and derive estimation and selection consistency, and Yoon, Park, and Lee (2013) who build on these results by estimating the autoregressive order directly from the data and by considering additional penalization methods. Hsu et al. (2008) derive the asymptotic theory for the lasso estimator under vector autoregressive (VAR) processes, and Kock (2016) considers application of the lasso to both stationary and nonstationary autoregressive processes.

Others have explored the realm of double-asymptotics, allowing the number of candidate variables to grow along with the sample size. Nardi and Rinaldo (2011) consider the estimation of autoregressive (AR) models where the number of lags increase with the sample size. Song and Bickel (2011) consider the (group-)lasso to estimate VAR models where the number of candidate variables is allowed to increase, but the number of relevant variables is kept fixed. Kock and Callot (2015) also use the lasso for VAR estimation, while allowing the number of relevant variables to increase. They provide non-asymptotic bounds and sufficient conditions for asymptotic consistency of the predictions, parameter estimates and variable selection. Unfortunately the generality of their results comes at the cost of imposing independence and normality on the errors. Medeiros and Mendes (2016) show that the adaptive lasso estimator maintains its consistency under substantially weaker assumptions and that the estimates are asymptotically normal even under weakly dependent errors. These results hold for (conditionally) heteroskedastic processes as well, although efficiency gains can be made through the use of alternative weighting (e.g. Wagener & Dette, 2013; Ziel, 2016). Thus, research has progressed to a point where lasso-type estimators are theoretically justifiable in a time series context and the applied econometrician is now required to choose between two appealing, though rather contrasting, approaches to modelling high-dimensional data.

Tuning

The implementation of lasso-type estimators requires the user to provide an a priori choice on the tuning parameters (λ, α). In the simulation exercises and the empirical application to follow, the tuning parameters are determined by obtaining the solution to (3) on a (100×1) grid of λ-values for the methods with a pre-determined α value or a (100×6) dimensional grid with (λ, α)-tuples for the (adaptive) elastic-net. We then use an information criterion, BIC or AIC, or time series cross-validation (CV) to select the optimal value(s). Time series CV is performed by reserving the first part of the sample to estimate the model under various settings of the tuning parameters after which the resulting models' fit are compared in a pseudo out-of-sample evaluation (Hyndman & Athanasopoulos, 2014). To illustrate, we adopt the threshold $c_T = \lceil \frac{2}{3} \times T \rceil$ and let $\boldsymbol{Z}_{c_T} = (\boldsymbol{W}_{c_T}, \boldsymbol{X}_{c_T})$, where $\boldsymbol{W}_{c_T} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{c_T})'$ and $\boldsymbol{X}_{c_T} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{c_T})'$. For a given value of the tuning parameter, say λ_j for $j \in J = \{1, \ldots, 100\}$, the model is estimated on $\boldsymbol{Z}_{c_T}$ to obtain the coefficient vector $\hat{\boldsymbol{\beta}}(\lambda_j)$. Next, a pseudo out-of-sample mean squared forecast error is calculated as $MSFE(\lambda_j) = \frac{1}{T-c_T} \sum_{t=c_T+1}^{T} (y_{t+h} - \boldsymbol{z}_t' \hat{\boldsymbol{\beta}}(\lambda_j))^2$. This procedure is executed for all values of the tuning parameter in the predefined grid and the final tuning parameter is chosen as

$$\hat{\lambda} = \arg\min_{\lambda_j} MSFE(\lambda_j).$$

In time series settings, this method is often preferred over traditional k-fold CV, because the time structure of the data is kept intact.[2]

2.2. Factor models

The literature on factor models is vast, their use being motivated through the conceptualization of factors as unobserved and possibly dynamic processes related to the state of the economy that drive a large set of observed economic time series. Factor models attempt to summarize the candidate set $\boldsymbol{X}$ by a smaller number of factors and, in the dynamic case, their lagged realizations. In this factor framework, the variables in the candidate set admit the following representation

$$\boldsymbol{x}_t = \boldsymbol{\Lambda}(L)\boldsymbol{f}_t + \boldsymbol{e}_t, \tag{4}$$

where $\boldsymbol{\Lambda}(L) = (\boldsymbol{\lambda}_1(L), \ldots, \boldsymbol{\lambda}_N(L))'$, $\boldsymbol{\lambda}_i(L) = (\lambda_{i,1}(L), \ldots, \lambda_{i,s}(L))'$ and $\lambda_{i,j}(L)$ is a lag polynomial of possibly infinite order describing how variable i loads onto the dynamic factor j. The symbol $\boldsymbol{f}_t$ refers to an $(s \times 1)$ vector containing the common factors and $\boldsymbol{e}_t$ is a vector of idiosyncratic disturbances.

The majority of the literature on forecasting with factor models has, either explicitly or implicitly, relied on the assumption of finiteness of the lag polynomials $\lambda_{i,j}(L)$. This assumption allows the model to be cast in a static form with the representation

$$\boldsymbol{x}_t = \boldsymbol{\Lambda}\boldsymbol{F}_t + \boldsymbol{e}_t. \tag{5}$$

where $\boldsymbol{\Lambda}$ contains the coefficients in $\boldsymbol{\Lambda}(L)$, $\boldsymbol{F}_t = (\boldsymbol{f}_t', \ldots, \boldsymbol{f}_{t-q}')'$ is a vector of size r with $s \leq r \leq (q+1)s$ and $\boldsymbol{e}_t = (e_{1t}, \ldots, e_{Nt})'$. The extension to the purpose of forecasting our generic model (1) follows naturally by substituting the candidate variables for their factor representation:

$$\begin{aligned} y_{t+h} &= \boldsymbol{w}_t' \boldsymbol{\beta}_w + \boldsymbol{x}_t' \boldsymbol{\beta}_x + \epsilon_{t+h} \\ &= \boldsymbol{w}_t' \boldsymbol{\beta}_w + \boldsymbol{F}_t' \boldsymbol{\Lambda}' \boldsymbol{\beta}_x + \boldsymbol{e}_t' \boldsymbol{\beta}_x + \epsilon_{t+h} \\ &= \boldsymbol{w}_t' \boldsymbol{\beta}_w + \boldsymbol{F}_t' \boldsymbol{\beta}_f + u_{t+h}, \end{aligned} \tag{6}$$

with $\boldsymbol{\beta}_f = \boldsymbol{\Lambda}' \boldsymbol{\beta}_x$ and u_{t+h} being the composite error that includes the innovation ϵ_{t+h} and the loss of information from summarizing the data $\boldsymbol{e}_t' \boldsymbol{\beta}_x$. The reduction in dimension from N to r allows this model to be estimated with OLS and the dependent variable to be forecast as $\hat{y}_{T+h|T} = \boldsymbol{w}_T' \hat{\boldsymbol{\beta}}_w + \hat{\boldsymbol{F}}_T' \hat{\boldsymbol{\beta}}_f$. Estimating the factors $\hat{\boldsymbol{F}}_T$ can be done with a wide variety of algorithms, the most common of which we discuss next.

The method of principal components (PC) is a popular means of extracting static factors. For any given k, which need not be equal to the true number of static factors r, the standard method of principal components (PC) obtains a $(T \times k)$ matrix of factor estimates and a $(N \times k)$ matrix of estimated loadings by solving the objective function

$$\left(\hat{\boldsymbol{\Lambda}}^k, \hat{\boldsymbol{F}}^k\right) = \arg\min_{\boldsymbol{\Lambda}^k, \boldsymbol{F}^k} \sum_t (\boldsymbol{x}_t - \boldsymbol{\Lambda}^k \boldsymbol{F}_t^k)' \boldsymbol{\Omega}^{-1} (\boldsymbol{x}_t - \boldsymbol{\Lambda}^k \boldsymbol{F}^k{}_t) \tag{7}$$

with $\boldsymbol{\Omega} = \boldsymbol{I}_N$ and subject to the normalization $\boldsymbol{\Lambda}^{k\prime} \boldsymbol{\Lambda}^k / N = \boldsymbol{I}_k$ and $\boldsymbol{F}^{k\prime} \boldsymbol{F}^k$ being diagonal.

A drawback of forecasting with standard PC is that the quality of the estimated components that serve as inputs for the forecasting equation strongly depends on the structure inherent to the original data. For example, Boivin and Ng (2006) demonstrate that cross-sectional correlation in the idiosyncratic component of (5) is highly detrimental to the quality of the component estimates. In search for a more robust form of component estimation, they propose the use of weighted principal components (WPC) by replacing the unobserved inverted population covariance matrix $\boldsymbol{\Omega}^{-1}$ in (7) with a feasible estimate $\hat{\boldsymbol{\Omega}}^{-1}$. Boivin and Ng (2006, p. 185) propose several weighting rules to obtain feasible estimates such as their weighting "rule SWa", where $\hat{\boldsymbol{\Omega}}^{-1}$ is diagonal with the ith diagonal element equal to $\left(\frac{1}{T} \sum_{t=1}^{T} \hat{\boldsymbol{e}}_t \hat{\boldsymbol{e}}_t'\right)_{ii}^{-1}$. We explore the additional rules "SWb", "Rule1" and "Rule2" proposed in their original paper as well and refer to them by their original names respectively.

Another cited disadvantage of principal component analysis is that every component is a linear combination of all variables, while a common empirical observation is that for any given component large groups of variables may carry small, non-zero loadings (e.g. Croux & Exterkate, 2011; Stock & Watson, 2002b). Similar to the premise underlying the lasso, it may be favourable to estimate factors that depend only on a subset of the variables to reduce forecast variability and, when of interest, improve interpretability of the model. The solution brought forward in

[2] While standard k-fold CV is valid for purely autoregressive models with uncorrelated errors (Bergmeir, Hyndman, and Koo, 2018), we observe time series CV to perform similarly in the simulations and superior in the empirical application.

the literature takes the form of sparse principal component (SPC), variants of which occur in Jolliffe, Trendafilov, and Uddin (2003), Shen and Huang (2008) and Zou, Hastie, and Tibshirani (2006). More recently, Kristensen (2017) considers the use of SPC for macroeconomic forecasting and shows that, under suitable restrictions on the amount of shrinkage, the SPC estimator is consistent under assumptions similar to those in Stock and Watson (2002a). While no additional assumption on the sparseness of the loadings is required for its consistency, the use of SPC implicitly assumes that a sparse representation is most suitable from the perspective of the classical bias/variance tradeoff. In this paper we adopt the computationally beneficial approach of Shen and Huang to estimate the sparse principal components and refer the reader to their original paper for details.

An alternative method of imposing sparsity is proposed by Bai and Ng (2008) who argue for forecasting with factor-augmented regressions by applying principal components to a subset of the predictors selected with the use of shrinkage estimators such as the lasso. Given the intuitive appeal of this approach and the documented improvement in performance by Bai and Ng, we include their *LA*(*PC*)-approach by applying the lasso for the purpose of subset selection in the first stage and extracting factors from that subset using standard PC in the second stage.[3]

Rather than casting the dynamic factor model (4) in the static framework (5), one may want to estimate the dynamic specification directly. Forni, Hallin, Lippi, and Reichlin (2000) propose a method to directly estimate (4) by obtaining the s dynamic factors on the basis of a consistent estimate of the population spectral density matrix. However, since the recovery of the dynamic factor relies on the estimation of a two-sided truncated filter, this approach does not work well for forecasting at the end of the sample. Accordingly, Forni et al. (2005) propose an alternative approach that decomposes the long run variance of the candidate set into contributions by the common and idiosyncratic component and estimates the factor loadings such that the share of the long rung variance attributable to the common component is maximized. This method is henceforth referred to as FHLR.

An alternative approach of explicitly modelling the dynamics in the factor model is provided by the method of maximum likelihood. While the idea of estimating static factors by maximum likelihood date back to the early work of Chamberlain and Rothschild (1983), more recently Doz, Giannone, and Reichlin (2011) and Doz et al. (2012) have provided the theory for maximum likelihood estimation of factor models under much less restrictive assumptions on the dynamic structure of the factors and the idiosyncratic component. The model is estimated under a relatively strict set of assumptions, e.g. a diagonal covariance matrix of the idiosyncratic component, with the use of the Kalman smoother. Doz et al. then show that certain deviations away from the assumptions under which the estimates are obtained are asymptotically negligible, thus justifying the method for a much broader class of data generating processes. This method will henceforth be referred to as DGR.

Finally, in recent contributions Forni, Hallin, Lippi, and Zaffaroni (2015); Forni et al. (2016) develop a method to obtain estimates of the dynamic components without imposing finiteness on the factor space. Under general assumptions, the authors derive one-sided representation of the dynamic factor model that can be estimated and used for forecasting. Throughout the paper we will refer to this method of forecasting as FHLZ, while referring the interested reader to the cited papers for details.

Tuning

All of the methods described above require an a priori choice for the number of factors. As such, much attention has been given to the development of data driven criteria that may aid the researcher in this choice absent of knowledge of the true number of factors. The reference criteria for static factor models in most contributions are those provided by Bai and Ng (2002), whom propose two classes of information criteria that minimize the variance of the idiosyncratic component subject to a penalty depending on both N and T. This method, however, is often documented to overestimate or underestimate the true number of factor (e.g. Forni, Giannone, Lippi, & Reichlin, 2009), on the grounds of which we employ several alternative criteria in the comparisons to follow. We consider methods that use the same type of information criteria with an extra tuning parameter (Alessi, Barigozzi, & Capasso, 2010) or that directly exploit the structure of the eigenvalues in the sample covariance matrix (Ahn & Horenstein, 2013; Onatski, 2010). For the dynamic factor models we employ the criteria of Hallin and Liška (2007) to select the number of dynamic components s. The DGR approach requires specification of the of autoregressive order of the dynamic factors. This is determined by obtaining initial estimates of the factors by principal components and fitting a VAR model on these estimates with the lag order being selected by the AIC. Finally, we implement the FHLZ method by randomly dividing the cross section of N time series in $\lfloor \frac{N}{q+1} \rfloor$ blocks on which we: (1) estimate VARs with their order determined by the AIC, (2) recover the dynamic components and (3) use these dynamic components and their lags to predict the dependent variable by an OLS projection.[4] This three-step process is repeated 50 times and the predictions are averaged over all iterations to remove the added noise from the cross-sectional sampling.

In the remainder of the paper we will stick to the convention of tabulating results only for the tuning method

[3] Others have also considered the reverse order, i.e. first extracting principal components from the data and then performing shrinkage on those components (e.g. Kim & Swanson, 2014; Stock & Watson, 2012). Yet another possibility is to apply shrinkage alongside factor estimation by sparsely estimating the idiosyncratic component (e.g. Hansen & Liao, 2016; Luciani, 2014). These approaches, however, are not pursued here as they are less related to the central questions examined in this paper and since their theoretical properties and empirical performance are well documented in the cited papers.

[4] To take into account the complete dynamic structure, predictions ought to be obtained by filtering the estimated factors as in Forni et al. (2016). However, we find that the direct OLS projection frequently outperforms the filtered predictions, especially for multi-step predictions in the empirical application, which motivates our choice of implementation.

that obtains the best performance on the factor model under consideration. Additional comments on the performance of other tuning methods are provided whenever deemed informative.

3. Simulation study

Our simulation study can broadly be categorized into three main sections, namely simulations on a DGP with (1) stationary observable variables with a sparse coefficient vector, (2) stationary common factors driving a large set of time series, and (3) non-stationary and cointegrated variables. In every category, we vary additional DGP characteristics such as the level of non-sphericity in the error, the number of common factors and the strength of the cointegration relationship.

Stationary observable variables

We generate the first set of DGPs as stationary processes where the dependent variable depends on five observable explanatory variables and a possibly autoregressive error term:

$$\begin{aligned} y_{t+1} &= \boldsymbol{x}_t'\boldsymbol{\beta}_x + \sqrt{\theta}\epsilon_{t+1} \\ (1-\alpha L)\epsilon_{t+1} &= v_{t+1} \end{aligned} \tag{8}$$

with $\boldsymbol{x}_t \sim \mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma}_N)$ and $v_{t+1} \sim \mathbb{N}(0, 1)$. Let $\boldsymbol{\iota}_5$ be a (5×1) vector of ones and $\mathbf{0}_{N-5}$ an $((N-5) \times 1)$ vector of zeros, then $\boldsymbol{\beta}_x = (\boldsymbol{\iota}_5', \mathbf{0}_{N-5}')'$. The population covariance matrix is generated as

$$\boldsymbol{\Sigma}_N = \begin{bmatrix} 1 & \cdots & \rho^{|j-i|} \\ \vdots & \ddots & \vdots \\ \rho^{|i-j|} & \cdots & 1 \end{bmatrix}$$

which allows for regulation of the degree of pairwise correlation between variable i and j by varying the single parameter ρ. In addition, we randomize the order of the newly generated variables prior to the construction of $\boldsymbol{y}$ in order to avoid a clustering of correlation in neighbouring variables. Furthermore, the signal-to-noise ratio is controlled by setting $\theta = \frac{1-\alpha^2}{10}\boldsymbol{\beta}_x'\boldsymbol{\Sigma}_N\boldsymbol{\beta}_x$, which keeps the population signal-to-noise ratio constant for changes in dimensionality of the model, as well as changes in the degree of serial correlation.

At every trial we generate $T = 100$ observations to which we apply all of the methods covered in Section 2. For the shrinkage estimators we generate the 1-step ahead forecast as $\hat{y}_{T+1|T} = \boldsymbol{x}_T'\hat{\boldsymbol{\beta}}_x$, whereas the predictions from factor models are obtained as $\hat{y}_{T+1|T} = \hat{\boldsymbol{F}}_T'\hat{\boldsymbol{\beta}}_F$. This procedure is repeated over $J = 1000$ trials and we evaluate the forecast performance of model i by the mean squared forecast error (MSFE)

$$MSFE_i = \frac{1}{J}\sum_{j=1}^{J}(y_{j,T+1} - \hat{y}^i_{j,T+1|T})^2. \tag{9}$$

The MSFE is reported relative to the MSFE of the optimal, though infeasible, OLS oracle method which forecasts the dependent variable by applying OLS to the five relevant variables only. As a measure of the estimation accuracy we calculate the mean squared error as

$$MSE_i = \frac{1}{J}\sum_{j=1}^{J}\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^i_j\right\|_2^2, \tag{10}$$

and, again, report the MSE relative to the OLS oracle procedure. Given the misspecified nature of the factor models on the current set of DGPs, this metric is reported for the shrinkage estimators only.

The selection performance of the shrinkage estimators is evaluated according to two standard metrics; the metric *consistent* depicts the fraction of trials in which the shrinkage estimators exactly identify the sparsity pattern by selecting the five relevant variables only, whereas *conservative* depicts the fraction of trials in which at least all five relevant variables are included. Finally, we also report the average number of variables included by each method as #*variables*. Detailed results regarding the shrinkage estimators are gathered in Tables 1–2. The performance of the factor models is tabulated in Table 3.

The results in Table 1 emphasize the effect of changes in dimensionality by leaving out any cross-sectional and serial correlation ($\rho = \alpha = 0$). Panel A reports results for the low-dimensional case ($N = 10$). In terms of the mean squared forecast error penalized regression performs at least as well as OLS, with the exception of ridge regression. The latter is unsurprising given that ridge regression does not impose sparsity and is a biased estimator that aims to reduce the MSE through a favourable bias–variance trade-off. The ability to do so, however, hinges on the presence of multi-collinearity, which is not an issue in the current set-up. Focussing on the lasso-type methods, we observe that the forecast performance of the adaptively weighted variants is superior to their non-weighted counterparts and, with RMSFEs of 1.01, is comparable to the infeasible oracle estimator. Concerning the selection performance, three results stand out. First, selection of the tuning parameter(s) by the BIC seems to lead more frequently to exact identification of the five relevant explanatory variables compared to cross-validation. Second, an adaptive weighting of the tuning parameter substantially improves the consistent selection scores and results in smaller models on average. Third, all methods considered are able to include the five relevant variables in all trials.

While promising, the results so far are derived in a low-dimensional setting where the gain relative to traditional OLS is small and the often cited "curse of dimensionality" is far from an issue. Accordingly, panel B–C display the performance for $N = 50$ and $N = 100$. The relative forecasting performance of OLS and ridge regression deteriorates and the difference in RMSFE with the sparsity inducing methods becomes more pronounced, despite the unreported MSFEs of the latter methods increasing along with the dimensionality as well. The detrimental effects of an increase in dimensionality are perhaps most apparent in the selection performance, with exact identification of the sparsity pattern occurring at substantially lower frequencies. Given that the conservative selection remains 100%, the drop in consistent selection necessarily stems from the inclusion of additional irrelevant variables, most likely due

Table 1
Stationary observed variables: the effect of dimensionality.

	OLS	ridge		las		adalas		en		adaen	
		BIC	CV	BIC	CV	BIC	CV	BIC	CV	BIC	CV
Panel A: $N = 10$											
RMSFE	1.05	1.11	1.13	1.08	1.08	**1.01**	1.05	1.08	1.08	1.01	1.05
RMSE	2.13	2.47	2.91	2.07	2.35	1.21	1.84	2.07	2.46	1.21	1.95
consistent	0%	0%	0%	27%	13%	84%	52%	27%	11%	84%	35%
conservative	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
#variables	10.00	10.00	10.00	6.45	7.91	5.21	6.10	6.45	8.10	5.21	6.89
Panel B: $N = 50$											
RMSFE	1.92	1.75	1.85	1.20	1.20	**1.04**	1.12	1.20	1.21	1.04	1.13
RMSE	19.09	16.15	17.91	5.05	4.74	1.65	3.42	5.06	4.81	1.65	3.95
consistent	0%	0%	0%	12%	3%	60%	23%	12%	3%	60%	15%
conservative	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
#variables	50.00	50.00	50.00	8.31	15.69	5.85	11.98	8.32	15.82	5.85	16.42
Panel C: $N = 100$											
RMSFE	–	–	7.78	1.28	1.24	**1.08**	1.09	1.28	1.24	1.08	1.10
RMSE	–	–	139.42	6.85	5.90	2.69	3.01	6.85	5.96	2.67	3.25
consistent	–	–	0%	8%	3%	33%	15%	8%	3%	33%	12%
conservative	–	–	100%	100%	100%	100%	100%	100%	100%	100%	100%
#variables	–	–	100.00	9.75	19.47	6.56	10.51	9.76	19.70	6.58	11.04

Notes: Numerical entries in this table are averages obtained over 1,000 simulations relative to the OLS oracle method for all evaluation metrics described in Section 3. Results are given for the low, mid and high-dimensional case in panel A, B and C respectively.

Table 2
Stationary observed variables: the effect of correlation.

ρ	α	OLS	ridge		las		adaLas		en		adaen	
			BIC	CV	BIC	CV	BIC	CV	BIC	CV	BIC	CV
Panel A: RMSFE												
0.0	0.0	1.92	1.75	1.85	1.20	1.20	**1.04**	1.12	1.20	1.21	1.04	1.13
0.6	0.0	1.94	1.52	1.56	1.12	1.16	1.02	1.12	1.12	1.18	**1.02**	1.14
0.6	0.6	1.88	1.49	1.51	1.13	1.14	1.03	1.09	1.13	1.15	**1.03**	1.11
Panel B: Consistent												
0.0	0.0	0%	0%	0%	12%	3%	60%	23%	12%	3%	60%	15%
0.6	0.0	0%	0%	0%	4%	2%	44%	16%	4%	2%	44%	11%
0.6	0.6	0%	0%	0%	4%	2%	48%	16%	4%	2%	48%	11%
Panel C: Conservative												
0.0	0.0	50.00	50.00	50.00	8.31	15.69	5.85	11.98	8.32	15.82	5.85	16.42
0.6	0.0	50.00	50.00	50.00	9.28	15.45	6.24	11.49	9.30	16.22	6.26	15.48
0.6	0.6	50.00	50.00	50.00	9.20	15.63	6.16	11.55	9.20	16.40	6.17	16.15

Notes: See notes in 1. The metrics considered are: (A) the RMSFE , (B) Consistent, and (C) the number of variables. Within each panel the different rows correspond to different settings of the degree of cross-sectional correlation (ρ) and serial correlation (α).

to randomly induced collinearity. Indeed, the increase in the number of variables selected in the higher dimensional settings supports this conjecture.

A well-known problem for the lasso is the presence of multi-collinearity in the data, especially between relevant and irrelevant variables, which can lead to inconsistencies in the selection of the correct variables (e.g. Zhao & Yu, 2006; Zou, 2006). As such, we examine the forecasting and selection performance under varying degrees of cross-sectional and serial correlation in Table 2, whilst keeping the dimension fixed at $N = 50$. Noteworthy is that while the MSFE increases for all methods when introducing a higher degree of cross-sectional correlation (unreported), the relative MSFE decreases for ridge regression and varies only marginally for the lasso-based regressions. The former finding is in line with the proclaimed benefits of ℓ_2-penalization under multi-collinearity, whereas the latter finding hints that the presence of cross-sectional correlation does not seem to affect the forecasting performance of lasso-type estimators more than OLS. Panel B clearly depicts the deterioration in selection performance after the introduction of cross-sectional correlation. While the unreported metric for conservative selection remains 100% for all methods, the consistent selection is strongly affected by the presence of cross-sectional correlation. In line with the aforementioned reasoning on the selection performance in high-dimensional settings, this implies that high levels of collinearity lead to larger models with irrelevant variables being erroneously included at higher frequencies. Finally, the method by which we scale the idiosyncratic noise term controls for the increased variance induced by serial correlation and, consequently, the introduction of serial correlation has little effect on the relative forecasting or selection performance.

Finally, in Table 3 we examine the predictive capabilities of factor models in the current framework. For each

Table 3
Stationary observed variables: factor models.

	PC	WPC				SPC	LA(PC)	FHLR	FHLZ	DGR
		SWa	SWb	Rule1	Rule2					
Panel A: $N = 50$, $\rho = 0$										
RMSFE	**9.06**	9.44	9.17	9.85	9.85	9.10	9.16	9.82	9.75	9.68
nvar	3.40	1.92	2.48	1.00	1.01	3.40	3.30	1.00	1.00	1.00
Panel B: $N = 50$, $\rho = 0.6$										
RMSFE	**2.57**	2.69	2.67	3.24	4.17	2.59	3.39	4.66	4.79	4.68
nvar	10.00	9.79	9.96	7.17	4.89	9.98	5.16	1.00	1.00	1.00

Notes: See notes in 1. Panel A lists results for a DGP with uncorrelated variables, whereas panel B lists results for a DGP allowing for a maximum population correlation of 0.6 between variables.

factor model, the results are reported for the factor selection method that delivers the best performance. Unsurprisingly, on a DGP absent of common components the factor models display inferior performance compared to the shrinkage estimators in Table 2. While the forecast accuracy worsens less when the variables in the dataset are correlated (Panel B) and when the information criterion selects a higher number of components, failure to include as many components as there are variables in the original dataset inevitably leads to a loss of information that negatively affects the forecasting performance. As a result, the *PC*-type criteria of Bai and Ng (2002) tend to deliver the best forecast accuracy here as they select more components on average. On the contrary, the dynamic factor models demonstrate relatively poor performance mainly as a result of the Hallin and Liška criterion selecting only a single dynamic factor in all simulation trials.

Stationary common factors

We next turn to the case where a small number of common factors drive a larger set of time series. The data-generating process contains an approximate factor structure and is a simplified version of the Stock and Watson (2002a) set-up recently employed by Kristensen (2017):

$$\begin{aligned} x_{it} &= \boldsymbol{\lambda}_i' \boldsymbol{f}_t + e_{it} \\ (1 - \alpha L) e_{it} &= (1 + \theta^2) v_{it} + \theta v_{i+1,t} + \theta v_{i-1,t} \end{aligned} \tag{11}$$

with $\boldsymbol{\lambda}_i, \boldsymbol{f}_t \overset{iid}{\sim} \mathbb{N}(\boldsymbol{0}, \boldsymbol{I}_r)$. The random variable $v_{i,t}$ drives the idiosyncratic component and is generated from a standard normal distribution. We impose sparsity in the loadings by setting a fraction τ of them equal to zero. While sparsity here simply refers to the presence of exact zero elements in the loadings, our approach of setting a fraction of all loadings equal to zero does not contradict the classic assumption of dense factor loadings, i.e. $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}/N \to \boldsymbol{I}_r$. As a result, even though the method of sparse principal components is expected to be more efficient here, the use of "non-sparse" factor models remains theoretically justifiable. The variable to forecast is generated as

$$y_t = \boldsymbol{f}_t' \boldsymbol{\beta}_f + \epsilon_t \tag{12}$$

where $\boldsymbol{\beta}_f$ is an $(r \times 1)$ vector of ones and ϵ_t is a standard normal error term. Recall that the shrinkage estimators attempt to forecast y_{T+1} as $\hat{y}_{T+1|T} = \boldsymbol{x}_t' \hat{\boldsymbol{\beta}}_x$, whereas the factor models use the extracted factors to construct the forecast $\hat{y}_{T+1|T} = \hat{\boldsymbol{f}}_t' \hat{\boldsymbol{\beta}}_{\hat{f}}$. Forecasting performance is measured on the basis of the MSFE relative to the factor-augmented regressions with the true number of factors calculated by standard PC. The two-step procedure calls for an additional metric measuring the estimation precision of the factor estimates in the first step. Following Doz et al. (2012) and Kristensen (2017), we report the trace R^2 as a measure to determine how well the estimated factors span the space of the true factors, calculated as

$$R_F^2 = \frac{Tr\left(\boldsymbol{F}'\hat{\boldsymbol{F}}(\hat{\boldsymbol{F}}'\hat{\boldsymbol{F}})^{-1}\hat{\boldsymbol{F}}'\boldsymbol{F}\right)}{Tr\left(\boldsymbol{F}'\boldsymbol{F}\right)}, \tag{13}$$

where $\hat{\boldsymbol{F}} = (\hat{\boldsymbol{f}}_1, \ldots, \hat{\boldsymbol{f}}_T)'$. While the shrinkage estimators obviously do not extract factors on the observed variables, the trace R^2 remains informative when interpreted as a measure of the accuracy with which the factor space is approximated by the subset of variables chosen by a given shrinkage estimator. Hence, for the shrinkage estimators we estimate

$$R_X^2 = \frac{Tr\left(\boldsymbol{F}'\boldsymbol{X}_S(\boldsymbol{X}_S'\boldsymbol{X}_S)^{-1}\boldsymbol{X}_S'\boldsymbol{F}\right)}{Tr\left(\boldsymbol{F}'\boldsymbol{F}\right)}, \tag{14}$$

where $\boldsymbol{X}_S$ denotes the subset of variables included by the method under consideration. The results for the set of DGPs with a single factor driving the time series are reported in Table 4 and for the case of four common factors in Table 5. To focus the comparison on differences between the factor extraction methods, rather than the factor selection methods, we report the results using the true number of factors only.[5]

Table 4 - panel A reveals that the factor models manage to slightly outperform the shrinkage estimators on a DGP where the population covariance matrix of the idiosyncratic component is diagonal, i.e. $\alpha = 0$ and $\theta = 0$. The trace R^2s are close to unity, which for the factor models implies accurate recovery of a rotation of the unobserved factor. For the shrinkage estimators, the high R^2s indicate that the limited number of variables chosen seems to be sufficient for a reasonable approximation of the factor space. This finding is in accordance with the proposition

[5] While the performance differentials between factor extraction methods remain qualitatively similar under the use of factor selection criteria, we do note the general finding that under strong forms of non-sphericity and a DGP with four latent factors all criteria tend to underestimate the true number of factors, with the exception of the *PC*-type criteria which heavily overestimate the true number of factors. All factor selection methods are more accurate under spherical idiosyncratic disturbance.

Table 4
DGP with one common factor.

	PC	WPC				SPC	LA(PC)	FHLR	FHLZ	DGR	las	adalas	en	adaen	ridge	ols
		SWa	SWb	Rule1	Rule2											
Panel A: $\alpha/\theta/\tau = 0/0/0$																
RMSFE	1.00	0.98	0.96	1.12	1.36	1.00	1.09	0.96	1.03	**0.96**	1.15	1.17	1.09	1.07	1.35	1.87
nvar	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	20.14	15.81	37.55	37.92	50.00	50.00
R^2	0.96	0.97	0.97	0.95	0.92	0.96	0.96	0.97	0.96	0.97	0.98	0.98	0.99	0.99	0.99	0.99
Panel B: $\alpha/\theta/\tau = 0/0/0.4$																
RMSFE	1.00	0.95	0.93	1.18	1.54	0.98	1.03	0.92	1.03	**0.92**	1.11	1.07	1.11	1.06	1.38	1.80
nvar	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	16.96	14.45	27.08	36.29	50.00	50.00
R^2	0.94	0.95	0.95	0.92	0.87	0.94	0.94	0.95	0.93	0.95	0.97	0.96	0.97	0.98	0.98	0.98
Panel B: $\alpha/\theta/\tau = 0.5/1/0.4$																
RMSFE	1.00	0.97	0.98	1.00	1.06	0.98	1.04	0.95	0.80	0.96	0.26	**0.26**	0.26	0.26	0.27	0.30
nvar	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	39.35	33.23	39.42	33.19	50.00	50.00
R^2	0.41	0.41	0.42	0.42	0.39	0.42	0.40	0.43	0.55	0.44	1.00	0.99	1.00	0.99	1.00	1.00

Notes: The reported RMSFEs are relative to the PC estimator that uses a single components in the forecasting equation. Each panel corresponds to a different setting of the degree of serial correlation (α), cross-sectional correlation (θ) and sparsity in the loadings (τ).

Table 5
DGP with four common factors.

	PC	WPC				SPC	LA(PC)	FHLR	FHLZ	DGR	las	adalas	en	adaen	ridge	ols
		SWa	SWb	Rule1	Rule2											
Panel A: $\alpha/\theta/\tau = 0/0/0$																
RMSFE	1.00	1.04	0.96	1.23	1.63	1.00	1.11	0.96	1.22	**0.96**	1.22	1.20	1.16	1.13	1.24	1.88
nvar	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	19.89	16.17	38.44	39.83	50.00	50.00
R^2	0.96	0.96	0.97	0.95	0.90	0.96	0.93	0.97	0.93	0.97	0.93	0.91	0.97	0.97	0.99	0.99
Panel B: $\alpha/\theta/\tau = 0/0/0.4$																
RMSFE	1.00	0.94	0.92	1.23	1.90	1.00	1.07	0.93	1.13	**0.92**	1.17	1.15	1.15	1.11	1.24	1.69
nvar	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	22.26	18.15	36.56	39.96	50.00	50.00
R^2	0.94	0.95	0.95	0.91	0.81	0.94	0.90	0.95	0.91	0.95	0.93	0.92	0.96	0.96	0.98	0.98
Panel B: $\alpha/\theta/\tau = 0.5/1/0.4$																
RMSFE	1.00	0.98	1.00	1.01	1.16	1.00	0.98	0.97	0.84	0.97	0.33	0.33	0.33	**0.33**	0.33	0.36
nvar	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	43.30	37.92	43.26	37.90	50.00	50.00
R^2	0.51	0.51	0.51	0.47	0.41	0.50	0.48	0.52	0.55	0.52	0.99	0.97	0.99	0.97	1.00	1.00

Notes: See notes in Table 4. The RMSFE is relative to the standard PC estimator that extracts four components.

of De Mol et al. (2008) who reason that the factor-induced collinearity in the candidate set allows for a few appropriately selected variables to capture the majority of the covariance in the data and to span approximately the same space as the common factors. Finally, ridge regression performs slightly worse than the lasso-type estimators and the OLS estimator displays the lowest forecast accuracy of all methods, despite obtaining the highest R^2. The latter finding can be considered as another example by which shrinkage estimators are able to improve upon the forecasting performance through a favourable bias–variance trade-off.

According to De Mol et al. (2008), forecasts from lasso-type estimators should not be expected to outperform correctly specified factor-augmented regressions, since the subset of the data proposed by methods employing an ℓ_1-penalty offers merely an approximation to the factor space and variable selection under high degrees of collinearity is known to be unstable. Indeed, panel B of Table 4 shows that the shrinkage estimators still underperform the factor models even when the component loadings are sparse. However, in panel C we observe that, after the introduction of substantial non-sphericity in the idiosyncratic component, the forecasting performance is tilted in favour of the shrinkage estimators. Under high levels of non-sphericity the factor models have difficulty in accurately estimating the unobserved factors, as indicated by the decrease in trace R^2s, whereas the shrinkage estimators tend to select a higher number of variables on average and, as a result, are able to maintain accurate approximation of the factor space. These patterns are similarly observed in the DGP with four factors, the results of which are displayed in Table 5, and provide a clear argument in favour of lasso-type estimation on data possessing factor structures with potentially non-spherical idiosyncratic components.

Upon further analysis, the introduction of cross-sectional correlation in the error term in (11) appears to be the main culprit for the deterioration in factor quality estimates. In the DGP with four factors, the percentage of the variance in the candidate set $\boldsymbol{X}$ explained by the first four standard estimated principal components is 72.3% before the introduction of cross-sectional correlation ($\alpha = 0.5, \theta = 0$) and 41.1% afterwards ($\alpha = 0.5, \theta = 1$). This is visualized in Fig. 1, where we display the ten largest eigenvalues of the sample correlation matrix corresponding to the first ten principal components. We conjecture that the correlation between the series in the candidate set that is induced by the idiosyncratic component obscures the

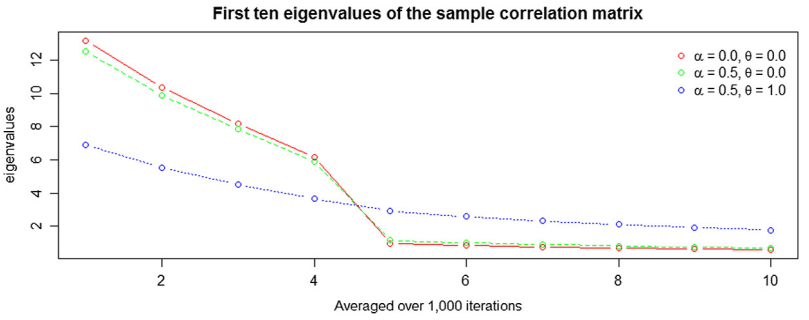

Fig. 1. Visualization of the explanatory power of the first ten common components.

factor-induced variation, thereby reducing the precision by which the factors are estimated, although we postpone a theoretical investigation on this phenomenon to future research.

Non-stationary and cointegrated variables

The presence and consequences of non-stationary predictors in regression frameworks are well-understood and numerous tests and solutions have been been proposed to correct for non-stationarity. Accordingly, in the majority of simulations and empirical work the implicit assumption is maintained that the researcher is able to successfully identify non-stationarity and all variables found to be integrated of order one or higher are transformed to stationarity by taking appropriate differences. However, situations are frequently encountered where the order of integration remains ambiguous (e.g. fractionally integrated variables or weakly cointegrated variables). In addition, the act of "correcting" for non-stationarity by differencing the variables comes at the cost of losing information captured in the levels of the variables. The literature on cointegration shows that long-run relationship between non-stationary variables can exist, relationships that are impossible to recover when using differenced variables. Here we examine the potential of lasso-type estimators in identifying and utilizing cointegrating relationships for forecasting in high-dimensional systems.

The potential for penalized regression in recognizing cointegrating relationships has recently been explored by Liang and Schienle (2015), Liao and Phillips (2015) and Wilms and Croux (2016) who all consider the use of penalized regression in automated vector error correction model estimation. These novel and insightful contributions, however, require a non-standard and fairly technical implementation. In an attempt to avoid placing this burden on the researcher, we focus on the use of an intuitive single equation model rather than a multivariate model. We generate the data as an error correction model:

$$\Delta y_t = \alpha \left(y_{t-1} - \sum_{i=1}^{3} \beta_i x_{i,t-1} \right) + \epsilon_{j,t} \tag{15}$$

$$x_{i,t} = x_{i,t-1} + \epsilon_{j+1,t} \qquad i = 1, 2, 3,\ j = 1, 2, 3$$

where the stationarity condition is given by $-2 < \alpha < 0$ and $\boldsymbol{\epsilon}_t \sim \mathbb{N}(\mathbf{0}, \boldsymbol{I}_4)$. In addition to the three variables $x_{i,t}$ for $i = 1, \ldots, 3$ that cointegrate with y_t we add a number of irrelevant variables to the candidate set $\boldsymbol{X}$. The high sample correlations induced by variables that are integrated of order one, i.e. $I(1)$, may have adverse consequences on the prediction and selection performance of the shrinkage estimators. Accordingly, we perform two sets of simulations; one in which the irrelevant variables are generated according to (8) with $\rho = 0.5$, $\alpha = 0$, and one in which half of the irrelevant variables are generated similarly, but the other half are generated as random walks, i.e. $\Delta x_{k,t} = \epsilon_{k,t}$ with $\epsilon_{k,t} \sim \mathbb{N}(0, 1)$. The two sets of simulations are simply referred to as "Stationary" and "Non-Stationary". As an example, for a candidate set $\boldsymbol{X}$ of size $N = 50$ that is generated in the Non-Stationary set, the first three variables will be $I(1)$ but cointegrated with the dependent variable. In the set of irrelevant variables, $\lceil \frac{N-3}{2} \rceil = 24$ are I(0) and $\lfloor \frac{N-3}{2} \rfloor = 23$ are $I(1)$. In congruence with the preceding simulations, we generate 1000 one-step ahead forecasts and report the metrics RMSFE and RMSE relative to the oracle OLS procedure as measures of prediction and selection performance respectively. The selection performance is, again, measured with the metrics *consistent*, *conservative* and *#variables*. The use of factor models is excluded from this section on the grounds that extracted factors can contain linear combinations of non-stationary variables and, hence, will be integrated of order one. Indeed, the presence of stochastic trends in the factors necessitates the use of alternative methods, such as the factor-augmented error correction model by Banerjee and Marcellino (2009), the forecasting performance of which is considered in Banerjee, Marcellino, and Masten (2014), or estimation of the factors in a VECM framework in the spirit of Barigozzi, Lippi, and Luciani (2016a, b). A preliminary analysis confirms that the factor models considered in this paper all display sub-par performance and are therefore omitted from the current analysis. We present the main results for the remaining estimators in Table 6, where the adjustment rate is fixed at $\alpha = -1$ and all tuning parameters are optimized based on the BIC. The effect of changes in the adjustment rate are further explored in Table 7.

Focussing on the predictive capabilities first, the RMSFEs in panel A of Table 6 demonstrate a superior performance of the ℓ_1 methods. The minimum RMSFE, denoted in bold, is always obtained by an adaptively weighted lasso-type estimator. Notwithstanding an overall decrease in forecasting performance relative to the OLS oracle procedure, the comparative advantage of lasso-type methods

Table 6
Cointegrated variables.

	Stationary			Non-Stationary		
	$N = 10$	$N = 50$	$N = 100$	$N = 10$	$N = 50$	$N = 100$
Panel A: RMSFE						
OLS	1.10	1.83	–	1.11	2.20	–
ridge	1.37	2.10	18.84	1.40	1.74	6.88
las	1.17	1.51	1.74	1.17	1.58	1.82
adalas	**1.03**	**1.09**	1.45	**1.05**	1.34	**1.60**
en	1.17	1.51	1.74	1.18	1.58	1.81
adaen	1.03	1.09	**1.43**	1.05	**1.34**	1.63
Panel B: RMSE						
OLS	9.38	106.70	–	7.48	89.98	–
ridge	9.89	64.72	46.26	11.61	51.82	46.61
las	4.22	8.21	10.64	5.31	18.88	26.90
adalas	2.16	3.25	8.37	2.51	16.39	24.86
en	4.22	8.20	10.78	5.33	18.98	27.10
adaen	2.16	3.24	8.08	2.52	16.46	25.14
Panel C: Consistent						
las	29.9%	20.1%	18.2%	9.8%	0.2%	0.0%
adalas	81.6%	62.4%	33.8%	63.8%	4.4%	0.2%
en	29.9%	20.0%	18.1%	9.9%	0.2%	0.0%
adaen	81.2%	62.2%	33.5%	63.6%	4.1%	0.2%
Panel D: Conservative						
las	99.5%	93.1%	88.5%	99.6%	82.5%	64.1%
adalas	99.8%	99.6%	91.2%	99.9%	79.3%	58.8%
en	99.5%	93.2%	88.5%	99.6%	82.3%	63.8%
adaen	99.8%	99.6%	91.6%	99.9%	79.3%	58.2%
Panel E: #Variables						
las	4.53	6.29	6.65	5.35	9.97	12.17
adalas	3.24	3.75	5.71	3.49	7.59	10.17
en	4.53	6.30	6.72	5.35	9.97	12.23
adaen	3.24	3.75	5.66	3.49	7.61	10.13

Notes: Numerical entries in this table are averages obtained over 1,000 simulations relative to the OLS oracle estimator that estimates the cointegrating vector with the cointegrated variables only. The methods considered are listed in the first column, whereas the evaluation metrics are divided across panels A–E. The results under "Stationary" are derived on a DGP absent of irrelevant I(1) variables, whereas those listed under "Non-Stationary" are derived on DGPs that do contain irrelevant I(1) variables.

relative to OLS or ridge becomes more pronounced for higher dimensions. The advantage of adaptive weighting over non-weighted estimation is substantial for the dimensions $N = 10$ and $N = 50$, but seems to diminish at $N = 100$. This most likely results from a deterioration in quality of the initial estimator, thereby highlighting the importance of finding good initial estimators in the high-dimensional setting. The estimation accuracy of the cointegrating vector, as measured by the RMSE, follows the same pattern as the prediction performance, with adaptively weighted estimation providing the highest accuracy and outperforming OLS even in the low-dimensional setting.

The selection performance is depicted in the remaining three panels of Table 6. Panel C depicts the fraction of trials in which the lasso-type methods identify the sparse cointegrating relationship exactly. Again, the adaptively weighted variants show superior performance. Exact identification, however, occurs at considerably lower rates in higher dimensional settings, with the decline in selection performance being most notable for the adaptively weighted estimators. A direct comparison between the scores for the consistent metric obtained on the Stationary and Non-Stationary sets reveals that the presence of irrelevant $I(1)$ variables negatively affects the selection performance. We conjecture that the inevitable high correlation between the non-stationary variables in levels, regardless of their relevance to the dependent variable, increases the difficulty in identifying the correct subset. Given that exact identification seems to be overly ambitious in this framework, we turn our attention to conservative selection. Absent of irrelevant non-stationary variables in the candidate set, the lasso-type methods almost always include at least all relevant variables. With the inclusion of additional $I(1)$ variables, we observe a worsening of the conservative selection, especially at higher dimensions, albeit not to levels as inadequate as observed for the consistent selection. Finally, the reason for conservative selection staying at reasonable levels can at least partly be attributed to the growing model size along increases in dimensionality. More irrelevant variables tend to be included when estimating on a larger candidate set and this effect is particularly apparent when non-stationary variables are present. Despite the faulty model selection characteristics

Table 7
Cointegrated variables: the effect of α.

	Stationary			Non-stationary		
	$\alpha = -1.9$	$\alpha = -1.0$	$\alpha = -0.1$	$\alpha = -1.9$	$\alpha = -1.0$	$\alpha = -0.1$
Panel A: Levels						
RMSFE	**1.21**	**1.13**	1.09	**1.34**	1.25	0.38
MSFE	25.77	4.68	16.33	30.15	5.53	5.58
Consistent	31.7%	57.3%	14.5%	16.8%	7.9%	0.0%
Conservative	79.1%	97.0%	32.3%	59.8%	89.0%	12.8%
Variables	4.00	3.95	3.00	4.42	6.86	12.66
Panel B: ADF differences						
RMSFE	3.54	2.14	0.14	3.48	1.73	0.14
MSFE	75.34	8.85	2.06	78.52	7.67	2.08
Consistent	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Conservative	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%
Variables	0.43	0.42	0.36	0.46	0.50	0.48
Panel C: Oracle differences						
RMSFE	3.64	1.21	**0.08**	3.58	**1.17**	**0.08**
MSFE	77.48	5.03	1.16	80.74	5.18	1.23
Consistent	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Conservative	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Variables	1.95	0.57	0.38	0.41	0.38	0.43

Notes: See notes in Table 6. The evaluation metrics considered are listed in the first column. The models are either estimated with all variables in levels (A), transformed variables based on the results of an ADF-test for stationarity (B) or infeasibly transformed variables based on knowledge of the true DGP (C).

in this non-stationary framework, the reduction in variance by excluding at least part of the irrelevant variables contributes enough to obtain a superior forecasting performance. Hence, for the applied researcher whose main interest lies in forecasting rather than model interpretation this somewhat naive application of lasso-type methods to cointegrated data in levels delivers substantial benefit.

The results so far are based on the somewhat idealized adjustment rate of $\alpha = -1$. If the adjustment rate would be closer to the lower boundary of the stationarity condition the dependent variable would show signs of negative autocorrelation that often characterizes an over-differenced time series, whereas a value close to the upper boundary would induce stronger dependence due to a slower adjustment rate. In both cases, the strength of the cointegrating relationship diminishes and a natural question that arises is how the lasso-type methods handle such situations. Furthermore, when the adjustment rate is small in magnitude, e.g. $\alpha = -0.1$, the equilibrium correction may be so slow that for the purpose of forecasting it is best to model the data in differences regardless. In the following analysis we focus on the use of the adaptive lasso on a candidate set consisting of 50 variables and examine the effect of changes in the adjustment rate on both the prediction and selection performance. For every adjustment rate, we examine the performance of the model estimated in three specifications; (1) all variables in the candidate set enter in levels, (2) some of the variables enter in differenced form based on the outcome of an Augmented Dickey-Fuller (ADF) test for stationarity of size 0.05, and (3) all variables that are simulated as $I(1)$ variables enter the model in differenced form. These models are listed in panel A, B and C of Table 7, respectively. The lowest RMSFE for a given adjustment rate across the three specification is denoted with bold font.

Models estimated in levels (panel A) only attain reasonable selection for an adjustment rate of $\alpha = -1$. Moving the adjustment rate towards the boundaries of the stationarity condition generally results in an increase in MSFE. However, different from the previous experiments, the strength of the adjustment rate also affects the OLS oracle estimator which serves as benchmark. A surprising finding is that the adaptive lasso does substantially better than the OLS oracle estimator when the adjustment rate is slow ($\alpha = -0.1$) and the candidate set contains irrelevant $I(1)$ variables. We expect that the inclusion of a large number of unrelated random walks allows for a better in-sample fit resulting in a lower forecast error; since the reported forecasts are single step forecast, the improved in-sample fit may favour the predictive performance of the resulting spurious models, because the combined effect of the corresponding random coefficients is unlikely to push the prediction of the dependent variable far from its realized value. However, this statistical artefact cannot be expected to carry through to forecasts over longer horizons as the trending behaviour of the $I(1)$ variables will cause the predictions to drift away from the realisations. Indeed, in unreported analyses we find that the predictive superiority of the adaptive lasso on weakly cointegrated variables relative to the OLS oracle procedure vanishes at a forecast horizon of 10 steps and keeps deteriorating for longer horizons, as one would expect to be the case for forecasts with spurious regressions.

The models estimated on transformed data based on ADF-tests in panel B all obtain substantially higher RMSFEs, unless the equilibrium correction is small ($\alpha = -0.1$). Upon closer inspection, however, it becomes apparent that for these cases the adaptive lasso hardly incorporates any variables from the dataset, but rather forecasts the dependent variable by its time series average. The low RMSFEs obtained by this simple strategy imply that the use of cointegration with a slow adjustment rate has limited relevance for short-term forecasting purposes. Furthermore,

for all adjustment rates the differenced models almost never contain all relevant variables. This provides an argument in favour of the use of ℓ_1-penalized estimation in levels over the traditional approach of pre-processing the data, especially on datasets characterized by a "strong" cointegrating relationship ($\alpha = 1$). Finally, the infeasible models based on an oracle differencing procedure in panel C perform similar to the ADF-differenced data.

In conclusion, the use of lasso-type estimators on a high-dimensional non-stationary dataset containing cointegrated variables provides forecast gains over the traditional approaches of using OLS on pre-processed data. A caveat to these results is that we rely on the underlying assumption of cointegration being present in the data. In practice, the uncertainty surrounding the validity of this assumption possibly affects the relative performance of the lasso-type methods. The interrelationship between verifying the presence of cointegration and forecast performance is practically relevant and we aim to pursue this topic in future research.

4. Empirical application

Complementing the simulation results, we perform an empirical application on a popular U.S. macroeconomic dataset. The dataset consists of 133 time series observed at a monthly frequency covering January 1959 to June 2015 and is obtained from the Fred-MD website.[6] In consideration of potentially adverse consequences stemming from uncertainty regarding the presence of cointegration in empirical datasets, we refrain from estimation in levels as considered in the previous section and correct all series for non-stationarity, which for the majority of series entails taking either log differences (e.g. real variables) or log second differences (e.g. price indices). Eight series are forecast, four of which are measures of real economic activity: real production income (RPI); total industrial production (IP); real manufacturing and trade sales (RMTS); and number of employees on non-agricultural payrolls (EMP). The remaining four series are price indices: the producer index for finished goods (PPI); the consumer price index (CPIA); the consumer price index less food (CPIUL); and the personal consumption expenditure implicit price deflator (PCEPI). These series, including their transformations, are similar to those frequently used in the seminal and contemporaneous forecasting literature (e.g. Kristensen, 2017; Ludvigson & Ng, 2009; Stock & Watson, 2002b).

The forecasts are generated as projections of an h-step-ahead variable y_{t+h}^h onto a set of variables observed up to time t that possibly includes lags of the dependent variable. As a benchmark, we consider a simple univariate AR model that obtains its forecasts by fitting the forecasting equation

$$y_{t+h}^h = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i+1} + \epsilon_{t+h}, \tag{16}$$

where y_{t+h}^h is defined appropriately according to the order of integration, see Stock and Watson (2002b) for details. The AR lag length p, for $p \in \{0, \ldots, 6\}$, is determined by the BIC criterion, as is the case for all following methods. The penalized regressions obtain the forecasts by fitting

$$y_{t+h}^h = \alpha + \boldsymbol{x}_t' \boldsymbol{\beta}_x + \sum_{i=1}^{p} \beta_i y_{t-i+1} + \epsilon_{t+h}, \tag{17}$$

where the tuning parameters λ, α are selected using either the BIC, AIC or time series cross-validation. The autoregressive lags enter the model unpenalized across all specifications, their selection thus being dependent on the use of the BIC criterion rather than the penalty induced shrinkage. Finally, forecasts based on static representations, i.e. all PC-type methods and the FHLR method, fit

$$y_{t+h}^h = \alpha + \hat{\boldsymbol{F}}_t' \boldsymbol{\beta}_F + \sum_{i=1}^{p} \beta_i y_{t-i+1} + \epsilon_{t+h}, \tag{18}$$

where the number of factors r is either kept fixed at five or determined by one of the information criteria of Bai and Ng (2002). Forecasts with the dynamic factor models FHLZ and DGR are based on

$$y_{t+h}^h = \alpha + \sum_{k=1}^{q} \hat{\boldsymbol{f}}_{t-k+1}' \boldsymbol{\beta}_{f,k} + \sum_{i=1}^{p} \beta_i y_{t-i+1} + \epsilon_{t+h}, \tag{19}$$

where $\hat{\boldsymbol{f}}_t$ is a s-dimensional vector of estimated dynamic factors. The number of lags of the factors that enter the forecast equation, $q \in \{0, \ldots, 6\}$, as well as the number of lags of the dependent variable are chosen by the BIC. We purposely do not forecast the target variable by iterated one-step ahead forecasts of the common and idiosyncratic components as is proposed in for example Forni et al. (2016), because the empirical performance of the iterated approach towards multi-step forecasts turned out to be highly inferior to the direct approach when forecasting the four price series. A similar finding is mentioned in Marcellino, Stock, and Watson (2006) who consider the same series and compare direct and iterated forecasts with autoregressive models. While the detrimental effects of using iterated forecasts are slightly mitigated when modelling the price series as being $I(1)$, the favourable performance for direct forecasts persists. Accordingly, we opt to model the price series as $I(2)$ and report the results for the direct forecasts only.

We simulate real-time forecasting by calculating pseudo out-of-sample forecasts at horizons $h = 1$ and $h = 12$. An initial in-sample period covering 10 years of monthly observations is used to estimate the models by which to obtain the first out-of-sample prediction. For each new prediction, we keep the length of the in-sample period fixed and move the estimation sample forward by one period, i.e. we adopt a rolling window approach. The model is re-estimated prior to each prediction, including tuning parameter optimization, lag length selection, shrinkage and factor estimation. The forecasting performance is reported as the mean squared forecast error relative to the benchmark AR model. The comparison of forecasts is established based on the computation of Model Confidence Sets (MCS), as proposed by Hansen, Lunde, and Nason (2011). We largely follow their original implementation with the $T_{R,\mathcal{M}}$-statistic and $\alpha = 0.25$. However, we do not adopt the moving-block bootstrap (MBB) procedure, given that

[6] https://research.stlouisfed.org/econ/mccracken/sel/.

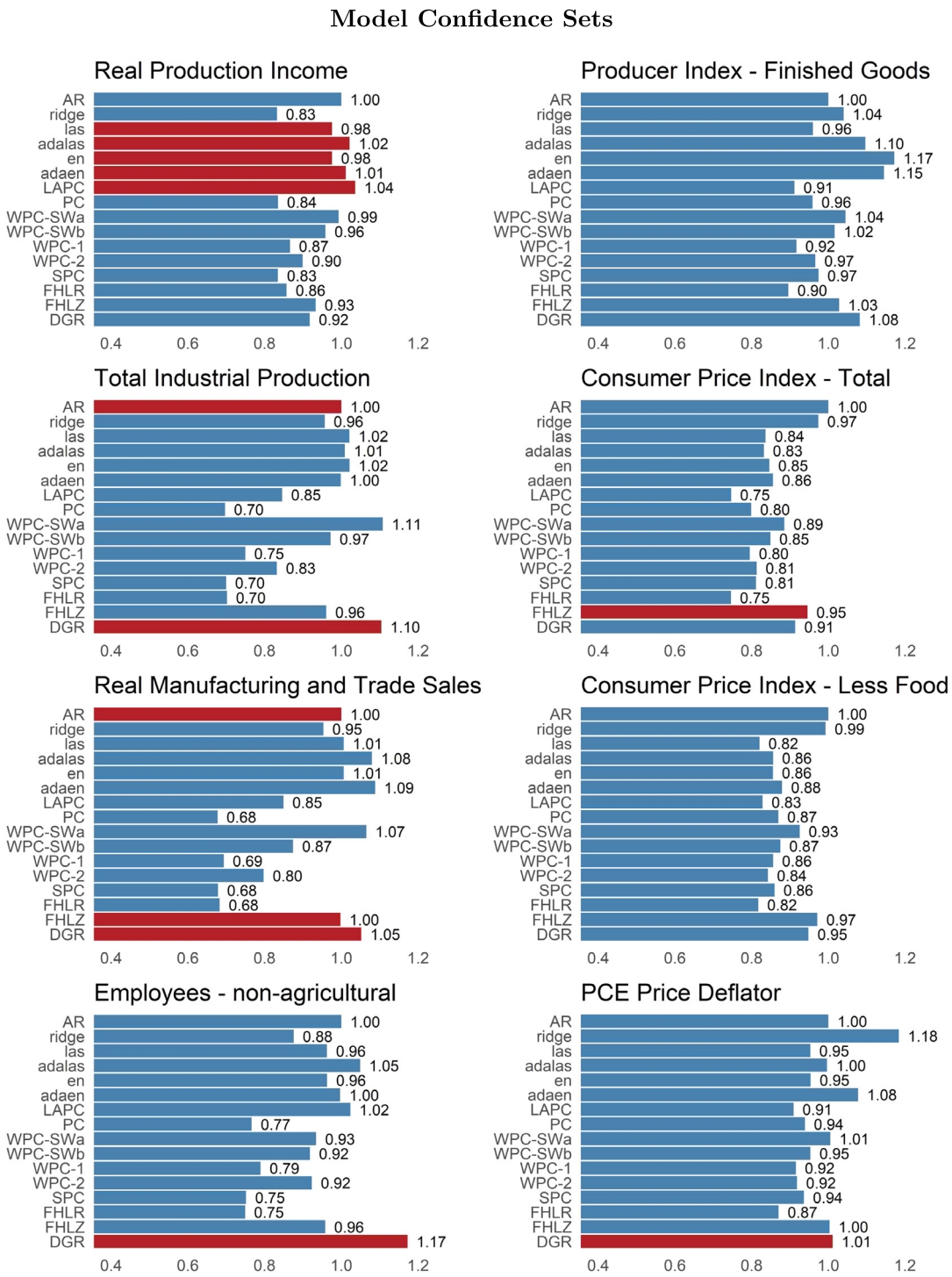

Fig. 2. Blue coloured bars represent members of the Model Confidence Sets. Results are for 12-month ahead forecasts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the time series of forecast errors display clear signs of unconditional heteroskedasticity over the full sample. Rather, we opt for the autoregressive wild bootstrap (AWB) which maintains its validity under the presence of both serial dependence and heteroskedasticity (Smeekes & Urbain, 2014). The autoregressive coefficient (γ) that governs the amount of dependence captured in the AWB is determined by fitting individual MA models to the series of forecast errors with their individual order being chosen by the AIC criterion. We use the median order of the MA models (q)

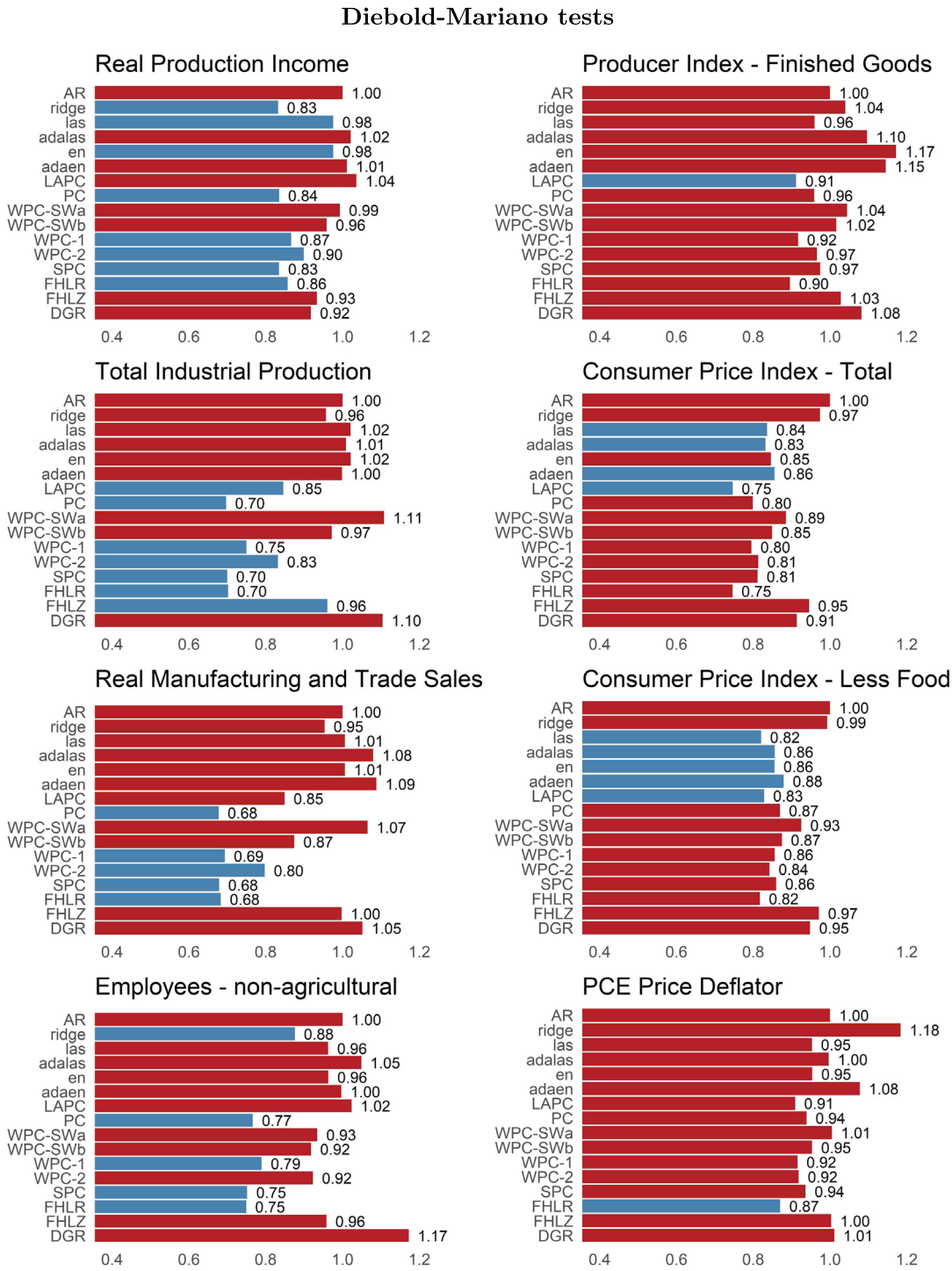

Fig. 3. Blue coloured bars represent models with RMSFEs significantly less than 1. Results are for 12-month ahead forecasts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

as a criterion for determining an appropriate block length, which we convert into the autoregressive coefficient with the conversion formula $\gamma = 0.01^{\frac{1}{q}}$ as proposed in Smeekes and Urbain (2014, p. 8). In a preliminary analysis, however, we find that the use of the MBB generally results in model confidence sets that contain the same models as those generated with the AWB.

We visualize the Model Confidence Sets graphically for the 12-month ahead forecasts in Fig. 2 while providing additional means of model comparisons with the use of

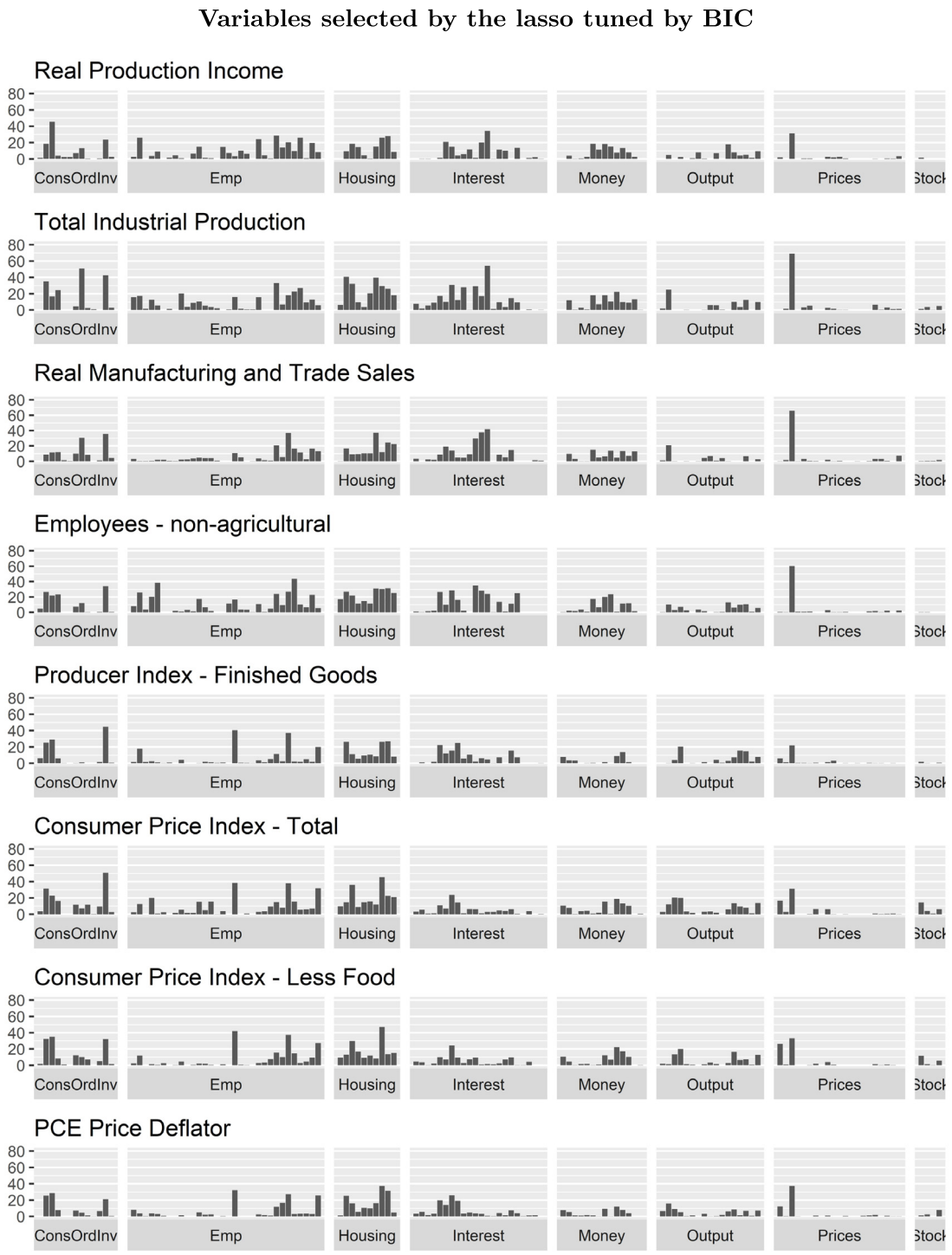

Fig. 4. The percentage of times a variable is included in the forecast equation, separated by economic category.

the Diebold–Mariano tests in Fig. 3. Comparisons of the monthly forecasts and a summary of the best performers are listed in the Appendix. The blue coloured bars in Fig. 3 represent the models contained in the MCS, while the red bars are removed and are thus considered to be models with statistically inferior predictive capability for the respective series-horizon. In absolute terms, we observe that for the real series (left column) the factor models seem to outperform the lasso-type methods with PC, SPC, and FHLR showing strong performance in particular, whereas the lasso-type methods are comparable to the factor models for the nominal series (right column). The

Fig. 5. An overview of the temporal selection properties per variable.

comparisons based on MCS almost always leaves all models in the set, seemingly suggesting that the variability in the forecast errors is too large to make any conclusive statements about the inferiority of certain models within the adopted 75% confidence level. The only exceptions to this are the exclusion of the lasso-type estimators for forecasts of Real Production Income (RPI) and occasionally some of the dynamic factor models FHLZ or DGR. An apparently counter-intuitive finding is that some of the methods removed from the MCS, e.g. the lasso in RPI, can have lower forecast losses than some of the models included in the MCS, e.g. "WPC-SWa" in RPI. The intuition behind

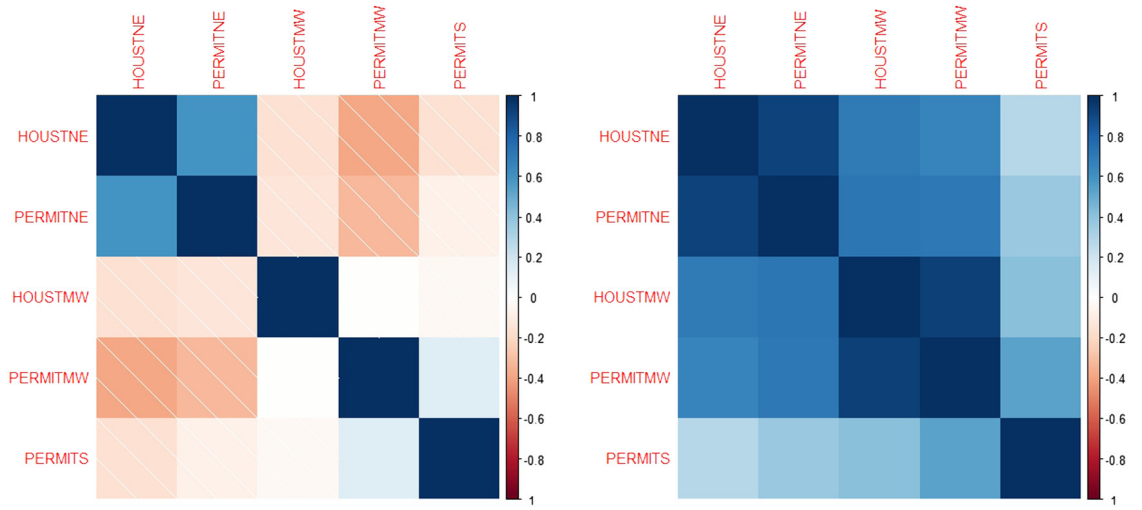

Fig. 6. Plots of correlations in the selection series (left) and absolute correlations in the realizations (right) of the housing series most frequently selected in "INDPRO" forecasts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

this curiosity is that the series that, despite their higher MSFEs, are included in the MCS display higher variability in their forecast errors which prevents one from concluding that the method performs worse than other methods with certainty, although one may rightfully wonder whether it is desirable to consider models with higher average loss superior simply because they display larger variation in their loss. Additionally, by controlling the familywise error rate (FWE), that is, the probability of making a single false rejection, the power of the MCS is highly dependent on the number of models considered. Our relatively large set of models is therefore detrimental in that respect. For this reason we also consider pairwise Diebold–Mariano tests which, by not controlling FWE, are not sensitive to this issue.

The Diebold–Mariano tests show frequent rejections of the null hypothesis of equal predictive capabilities in reference to the AR benchmark. The dominance of factor models on the real series and of the lasso-type estimators on the (consumer) price indices is immediately notable; on the real series most of the factor models are considered to obtain MSFEs significantly lower than the AR benchmark, whereas for the consumer price indices rejection only occurs for the methods involving L_1-shrinkage which is partially attributable to the lower variability in forecast errors of these methods. Finally, the dynamic factor methods FHLZ and DGR tend to perform slightly worse than the static variants, although we cautiously note that this may be a somewhat unfair comparison given the availability of a larger range of factor selection approaches for the static models. Indeed, during simulations we observed the Hallin and Liška criterion to occasionally deliver sub par performance. Given that the main comparison of interest, however, is the difference in predictive capability between shrinkage and factor methods we do not consider this caveat to impede our conclusions.

Hyperparameters and factor selection

We briefly comment on the performance of individual tuning methods for each model. The best performance by the shrinkage estimators is most frequently attained by tuning with the BIC criterion and CV coming in second place. For the static factor methods, the criteria most frequently leading to the best forecasting performance tend to be one of the Alessi et al. (2010) criteria, their IC3 criteria showing strong performance in particular. For the dynamic factor methods the use of a single dynamic factor performs best, followed by the use of four dynamic factors and the Hallin and Liška (2007) performs worst, possibly explaining the suboptimal predictive capability of the dynamic factor methods.[7] Lastly, the LA(PC) approach based on a preliminary lasso estimation performs similar when the lasso is tuned with either the BIC-criterion or the AIC-criterion.

Variable selection and sparsity patterns

The documented performance of the lasso-type methods may leave one wondering whether the assumption of latent factors driving the variation in observable economic time series is justified. We explore the proposition of De Mol et al. (2008) where the collinearity induced by latent factors allows for approximation of the factor space with relatively few observable variable, while simultaneously resulting in highly unstable variable selection. In Fig. 4 we display the fraction of 12-month ahead forecast equations in which each variable in the data is selected by the lasso tuned with the BIC criterion. Strikingly, the pattern of frequently chosen variables is fairly consistent across the different forecast series, in particular when considering the group of nominal and real target variables separately. For example, in the Prices category, the "ISM Manufacturing: Price Index" (NAPMPRI) seems to capture the majority of the variation, whereas for the housing category the variables seem to substitute each other based on

[7] We evaluate the Hallin and Liška criterion at three different sample points, i.e. (N_c, T_c) with $c \in \{1, 2, 3\}$, which is not necessarily optimal for the current empirical application.

Table 8
Most frequently selected variables.

Forecast	ConsOrdInv	Emp	Housing	Interest
RPI	NAPMSDI	USWTRADE	PERMITS	BAAFFM
INDPRO	BUSINVx	USWTRADE	HOUSTNE	BAAFFM
CMRMTSPLx	M2REAL	USFIRE	PERMITNE	BAAFFM
PAYEMS	M2REAL	USGOVT	PERMITS	T10YFFM
PPIFGS	M2REAL	CES1021000001	PERMITS	TB6SMFFM
CPIAUCSL	M2REAL	CES1021000001	PERMITMW	TB3SMFFM
CPIULFSL	NAPMSDI	CES1021000001	PERMITMW	TB3SMFFM
PCEPI	NAPMSDI	CES1021000001	PERMITMW	TB3SMFFM
Forecast	Money	Output	Prices	Stock
RPI	CONSPI	IPBUSEQ	NAPMPRI	DTCOLNVHFNM
INDPRO	S.P.PE.ratio	W875RX1	NAPMPRI	INVEST
CMRMTSPLx	CONSPI	W875RX1	NAPMPRI	INVEST
PAYEMS	S.P.div.yield	IPBUSEQ	NAPMPRI	DTCOLNVHFNM
PPIFGS	FEDFUNDS	CMRMTSPLx	NAPMPRI	INVEST
CPIAUCSL	S.P.PE.ratio	DPCERA3M...	NAPMPRI	INVEST
CPIULFSL	S.P.PE.ratio	CMRMTSPLx	NAPMPRI	INVEST
PCEPI	S.P.PE.ratio	W875RX1	NAPMPRI	INVEST

Notes: This table report the most frequently selected variables in 12-month ahead forecast by the lasso tuned with the BIC criterion. For an overview of all the variables and their abbreviations, see the appendix in McCracken and Ng (2015).

the low frequencies with which they are selected.[8] Not a single variable, however, is chosen consistently over all forecast periods. In line with the proposition of De Mol et al. (2008), this could be due to temporal instability resulting from collinearity induced by latent factors. Alternatively, structural changes may occur over the complete sample causing the relevance across variables to shift over time. To distinguish between these contrasting explanations we plot an overview of the variable selection over time in Fig. 5, where a green bar indicates that the variable was included in the forecast while a red bar indicates exclusion. The vertical axis contains the 515 12-month ahead forecasts performed. Directly observable is the persistence in the selection of the most frequently included variables in the consumption, employment and prices categories, for which the structural change explanation seems most applicable. For other categories, such as housing or interest, factor-induced collinearity may offer an appropriate description, however.

The housing category provides a particularly suitable subset to examine whether the overlap in informational content of individual time series allows for approximation of the factor space with only a few cleverly selected variables. We focus on the 12-month ahead forecasts of Total Industrial Production (INDPRO) and consider the five most frequently chosen housing variables. We construct five new binary time series that indicate whether a variable for a given forecast at time $t + h$ was included and we refer to these as the selection series. Under the conjecture that the selection is unstable because the individual variables approximate the same space, one would expect to observe negative correlation between the selection series due to substitution effects and this negative correlation between the selection series should be stronger for time series that exhibit strong correlation in their realizations. Accordingly, we list two correlation plots in Fig. 6. Evidence in favour of this conjecture would match up large negative correlation in the selection series, i.e. dark red boxes in the left plot, with large absolute correlations in the realizations of the respective series, i.e. dark blue boxes in the right plot. However, we observe that the selection series exhibit only mild negative correlation and the strongest correlated variables, i.e. "HOUSTNE" and "PERMITNE", actually tend to be selected together rather than substitute each other. We interpret these findings as anecdotal evidence that the variables selected by the lasso each contribute unique information and that structural change in the underlying DGP offers a feasible explanation of the temporal instability in the selection properties alongside the proposition of factor-induced collinearity in the observed time series.

5. Conclusion

In this paper we examine the forecasting performance of factor, shrinkage and hybrid models. Comprehensive simulations based on a wide variety of data generating processes indicate that lasso-type estimators are relatively robust against alternative DGP specifications; they naturally perform well on sparse and stationary models driven by observed variables, but they also show strong forecasting performance on data driven by approximate factor structures, even when the latter models contain a high degree of non-sphericity in the idiosyncratic component. Furthermore, a direct application of lasso-type estimators to a high-dimensional non-stationary dataset containing a small number of cointegrated variables is demonstrated to deliver forecasting improvements over traditional approaches. An empirical application on eight macroeconomic time series confirms the strong performance of factor-based model that is frequently covered in the forecasting literature. However, for certain target series such as the Consumer Price Index the lasso-type methods offer comparable if not better forecasting performance, while simultaneously displaying fairly persistent variable selection behaviour. We take this as further evidence that the assumption of common factors being persistent in macroeconomic data may not always be valid or, at a minimum,

[8] An overview of the most frequently chosen variable per economic category is provided in Table 8 in the Appendix.

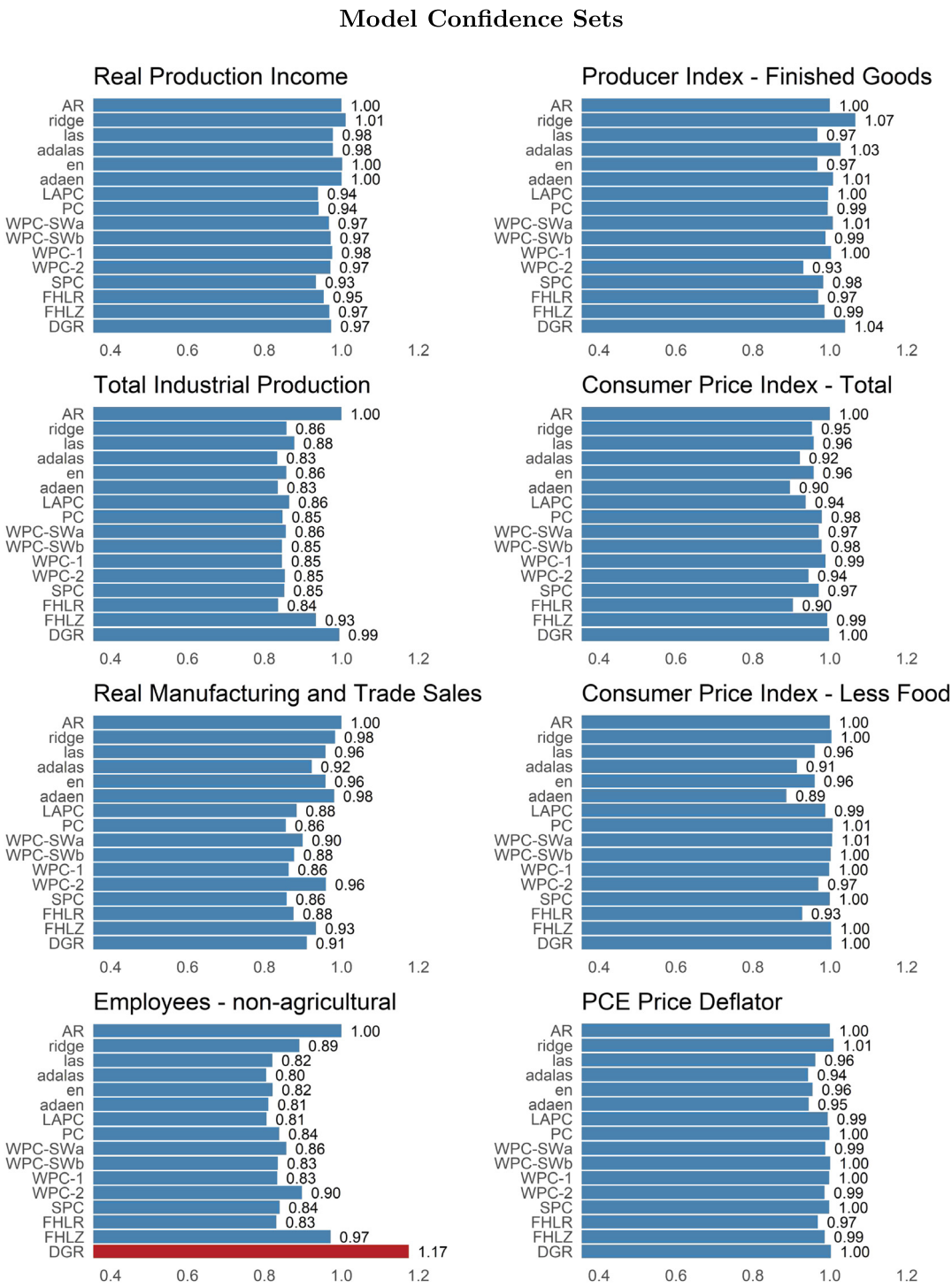

Fig. 7. Blue coloured bars represent members of the Model Confidence Sets. Results are for 1-month ahead forecasts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

may not always be relevant for forecasting purposes given the flexibility with which lasso-type estimators can handle this type of data.

Appendix. Empirical forecasts

See Figs. 7 and 8 and Table 8.

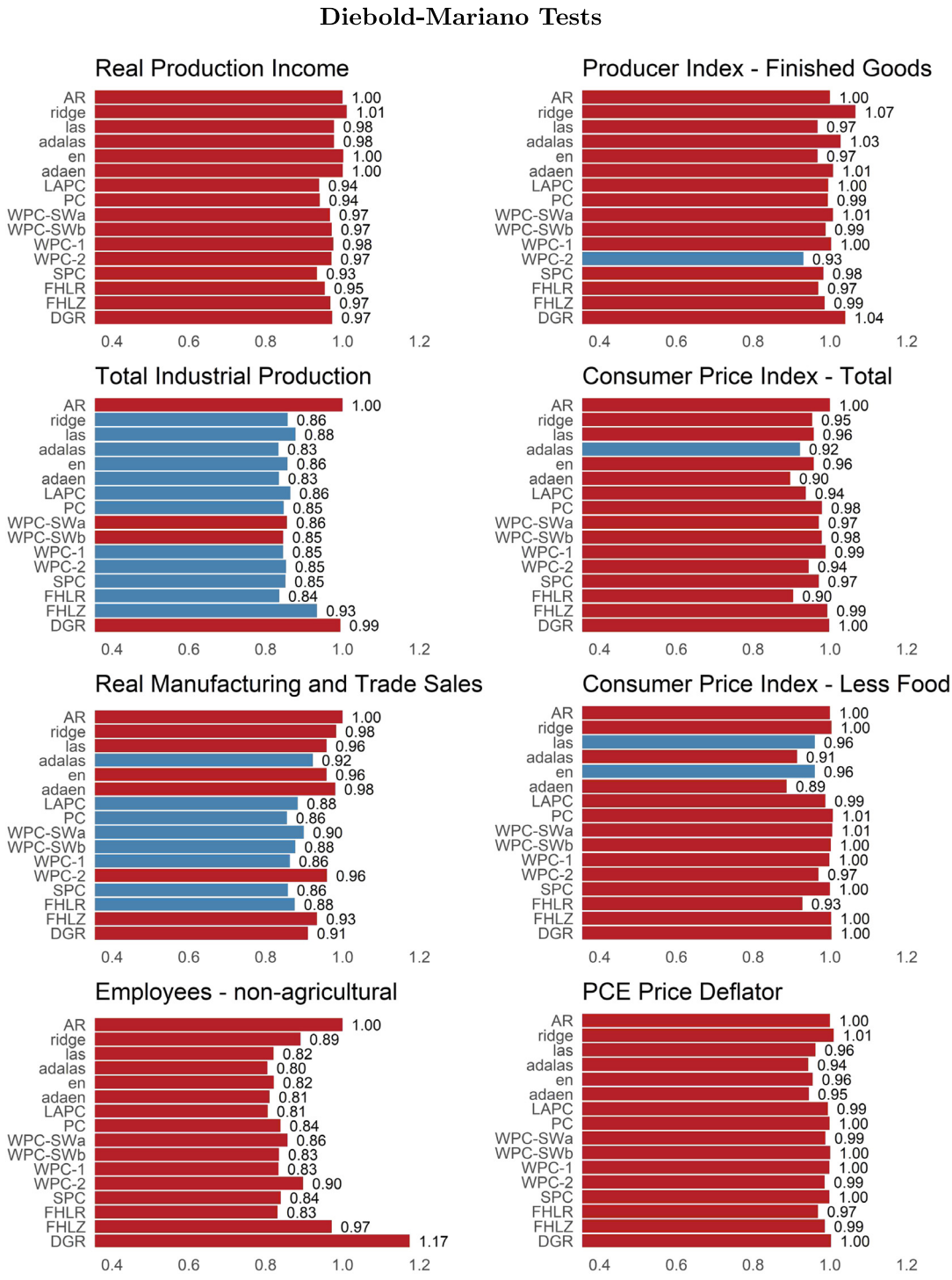

Fig. 8. Blue coloured bars represent models with RMSFEs significantly less than 1. Results are for 1-month ahead forecasts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

References

Ahn, S. C., & Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, *81*(3), 1203–1227.

Alessi, L., Barigozzi, M., & Capasso, M. (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics & Probability Letters*, *80*(23), 1806–1813.

Artis, M. J., Banerjee, A., & Marcellino, M. (2005). Factor forecasts for the UK. *Journal of Forecasting*, *24*, 279–298.

Bai, J., Li, K., & Lu, L. (2016). Estimation and inference of FAVAR models. *Journal of Business & Economic Statistics*, *34*, 620–641.

Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*, 191–221.

Bai, J., & Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, *146*, 304–317.

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

Banerjee, A., & Marcellino (2009). Factor-augmented error correction models. In J. L. Castle, & N. Shephard (Eds.), *The methodology and practice of econometrics —a festschrift for David Hendry* (pp. 589–612). Oxford: Oxford University Press.

Banerjee, A., Marcellino, M., & Masten, I. (2014). Forecasting with factor-augmented error correction models. *International Journal of Forecasting*, *30*(3), 589–612.

Barigozzi, M., & Brownlees, C. (2017). NETS: Network estimation for time series. https://ssrn.com/abstract=2249909.

Barigozzi, M., Lippi, M., & Luciani, M. (2016a). Dynamic factor models, cointegration, and error correction mechanisms. Working Paper. http://arxiv.org/abs/1510.02399.

Barigozzi, M., Lippi, M., & Luciani, M. (2016b). Non-stationary dynamic factor models for large datasets. Working Paper. http://ssrn.com/abstract=2741739.

Bergmeir, C., Hyndman, R. J., Koo, B., et al. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, *120*, 70–83.

Bernanke, B. S., Boivin, J., & Eliasz, P. (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics*, *120*(1), 387–422.

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

Boivin, J., Ng, S. (2006). Are more data always better for factor analysis? *Journal of Econometrics*, *132*, 169–194.

Callot, L. A., & Kock, A. B. (2014). Oracle efficient estimation and forecasting with the adaptive lasso and the adaptive group lasso in vector autoregressions. *Essays in Nonlinear Time Series Econometrics*, 238–268.

Chamberlain, G., & Rothschild, M. (1983). Factor structure, and mean-variance analysis on large asset markets. *Econometrica*, *51*, 1281–1304.

Croux, C., & Exterkate, P. (2011). Sparse and robust factor modelling. Tinbergen Institute Discussion Paper TI 122/4.

De Mol, C., Giannone, D., Reichlin, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, *146*, 318–328.

Doz, C., Giannone, D., & Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, *164*(1), 188–205.

Doz, C., Giannone, D., & Reichlin, L. (2012). A quasi–maximum likelihood approach for large, approximate dynamic factor models. *Review of Economics and Statistics*, *94*, 1014–1024.

Eickmeier, S., & Ziegler, C. (2008). How successful are dynamic factor models at forecasting output and inflation? A meta-analytic approach. *Journal of Forecasting*, *27*, 237–265.

Forni, M., Giannone, D., Lippi, M., & Reichlin, L. (2009). Opening the black box: Structural factor models with large cross sections. *Econometric Theory*, *25*(5), 1319–1347.

Forni, M., Giovannelli, A., Lippi, M., & Soccorsi, S. (2016). Dynamic factor model with infinite dimensional factor space: forecasting. https://ssrn.com/abstract=2766454.

Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2000). The generalized dynamic factor model: identification and estimation. *The Review of Economics and Statistics*, *82*, 540–554.

Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2005). The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association*, *100*(471), 830–840.

Forni, M., Hallin, M., Lippi, M., & Zaffaroni, P. (2015). Dynamic factor models with infinite-dimensional factor spaces: one-sided representations. *Journal of Econometrics*, *185*(2), 359–371.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22.

Gelper, S., & Croux, C. (2008). *Least angle regression for time series forecasting with many predictors*. KU Leuven-Faculty of Business and Economics.

Hallin, M., & Liška, R. (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*, *102*(478), 603–617.

Hansen, C., & Liao, Y. (2016). The factor-lasso and K-Step bootstrap approach for inference in high-dimensional economic applications. ArXiv Preprint arXiv:1611.09420.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497.

Hsu, N., Hung, H., & Chang, Y. (2008). Subset selection for vector autoregressive processes using lasso. *Computational Statistics & Data Analysis*, *52*, 3645–3657.

Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: principles and practice*. OTexts.

Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, *12*(3), 531–547.

Kascha, C., & Trenkler, C. (2015). *Forecasting VARs, model selection and shrinkage*. University of Mannheim / Department of Economics.

Kim, H. H., & Swanson, N. R. (2014). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics*, *178*, 352–367.

Knight, K., & Fu, W. (2000). Asymptotics for LASSO-type estimators. *The Annals of Statistics*, *28*, 1356–1378.

Kock, A. B. (2016). Consistent and conservative model selection with the adaptive lasso in stationary and nonstationary autoregressions. *Econometric Theory*, *32*, 243–259.

Kock, A. B., & Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, *186*, 325–344.

Kristensen, J. T. (2017). Diffusion indexes with sparse loadings. *Journal of Business & Economic Statistics*, *35*, 434–451.

Li, J., & Chen, W. (2014). Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models. *International Journal of Forecasting*, *30*, 995–1015.

Liang, C., & Schienle, M. (2015). *Determination of vector error correction models in higher dimensions*. Leibniz Universität Hannover.

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

Luciani, M. (2014). Forecasting with approximate dynamic factor models: The role of non-pervasive shocks. *International Journal of Forecasting*, *30*(1), 20–29.

Ludvigson, S. C., & Ng, S. (2009). A factor analysis of bond risk premia. *National Bureau of Economic Research*, *w15188*.

Marcellino, M., Stock, J. H., & Watson, M. W. (2003). Macroeconomic forecasting in the Euro area: Country specific versus area-wide information. *European Economic Review*, *47*, 1–18.

Marcellino, M., Stock, J. H., & Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, *135*(1), 499–526.

McCracken, M. W., & Ng, S. (2015). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics* (forthcoming).

Medeiros, M. C., & Mendes, E. F. (2016). ℓ_1-regularization of high-dimensional time series models with non-Gaussian and heteroskedastic errors. *Journal of Econometrics*, *191*, 255–271.

Nardi, Y., & Rinaldo, A. (2011). Autoregressive process modeling via the Lasso procedure. *Journal of Multivariate Analysis*, *102*, 529–549.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, *92*(4), 1004–1016.

Pesaran, M. H., Pick, A., & Timmerman, A. (2011). Variable selection, estimation and inference for multi-period forecasting problems. *Journal of Econometrics*, *164*, 173–187.

Shen, H., & Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, *99*, 1015–1034.

Smeekes, S., & Urbain, J. (2014). *A multivariate invariance principle for modified wild bootstrap methods with an application to unit root testing*. GSBE, Research Memorandum RM/14/008, Maastricht University.

Song, S., & Bickel, P. J. (2011). Large vector auto regressions. ArXiv Preprint arXiv:1106.3915.

Stock, J. H., & Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*, 1167–1179.
Stock, J. H., & Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, *20*, 147–162.
Stock, J. H., & Watson, M. W. (2006). Forecasting with many predictors. *Handbook of Economic Forecasting*, *1*, 515–554.
Stock, J. H., & Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, *30*, 481–493.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B.*, *58*, 267–288.
Wagener, J., & Dette, H. (2013). The adaptive lasso in high-dimensional sparse heteroscedastic models. *Mathematical Methods of Statistics*, *22*, 137–154.
Wang, H., Li, G., & Tsai, C. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, *69*, 63–78.
Wilms, I., & Croux, C. (2016). Forecasting using sparse cointegration. *International Journal of Forecasting*, *32*, 1256–1267.
Yoon, Y. J., Park, C., & Lee, T. (2013). Penalized regression models with autoregressive error terms. *Journal of Statistical Computation and Simulation*, *83*, 1756–1772.
Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.
Ziel, F. (2016). Iteratively reweighted adaptive lasso for conditional heteroscedastic time series with applications to ar-ARCH type processes. *Computational Statistics & Data Analysis*, *100*, 773–793.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.
Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*, 265–286.

Stephan Smeekes is Associate Professor at the Department of Quantitative Economics of Maastricht University. His research interests include time series econometrics and bootstrap methods. He has published articles in scientific journals such as *Journal of Econometrics*, *Econometric Theory*, *Journal of Business and Economic Statistics* and *Econometric Reviews*.

Etienne Wijler joined the Department of Quantitative Economics at Maastricht University in September 2015 as a doctoral candidate under the supervision of Stephan Smeekes and Jean-Pierre Urbain. His research interests include time series analysis, statistics for high-dimensional data and forecasting. Currently, the main focus of his research is on the use of penalized regression in non-stationary frameworks.