Macroeconomic Forecasting Using Penalized Regression Methods*

Stephan Smeekes Etiënne Wijler

Maastricht University
Department of Quantitative Economics
November 21, 2016

Abstract

We study the suitability of lasso-type penalized regression techniques when applied to macroeconomic forecasting with high-dimensional datasets. We consider performance of the lasso-type methods when the true DGP is a factor model, contradicting the sparsity assumption underlying 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 estimated by principal components, even if the underlying DGP is a factor model. Furthermore, the penalized regession methods are demonstrated to deliver forecast improvements over traditional approaches when applied to non-stationary data containing cointegrated variables, despite a deterioration of the selective capabilities. Finally, we also consider an empirical application to a large macroeconomic U.S. dataset and demonstrate that, in line with our simulations, penalized regression methods attain the best forecast accuracy most frequently.

Keywords: Forecasting, Lasso, Factor Models, High-Dimensional Data, Cointegration.
JEL-Codes: C22, C53, E17

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

*Department of Quantitative Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: S.Smeekes@maastrichtuniversity.nl, E.Wijler@maastrichtuniversity.nl. The first author would like to thank the Netherlands Organisation for Scientific Research (NWO) for financial support. We thank Alain Hecq, Rasmus Lönn, and Jean-Pierre Urbain for helpful discussions. We are grateful to conference and seminar participants at CFE-CMStatistics 2015 for helpful comments and suggestions. All remaining errors are our own.

stochastic trends and may be cointegrated. We also provide a systematic comparison with factor models estimated by principal components, the mainstream technique 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 PC-type estimators 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 and 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 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 PC-type estimators. 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. 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 Fred-MD database (McCracken and 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 al-

ternative DGP specifications. PC-type estimators 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 PC-type estimators on DGPs without factors, but more surprisingly also provides forecast improvements on DGPs with serially and cross-sectionally correlated factors. In addition, penalized regression shows promising results on cointegrated data, producing substantially lower forecasting errors compared to standard OLS despite failing to identify the exact cointegrating vector at relatively high frequencies. Finally, the empirical application is consistent with our findings on the simulated data, showing favourable performance of the shrinkage estimators across a variety of macroeconomic time series and forecast horizons.

Our contribution compliments the vast existing macroeconomic forecasting literature that is dominated by methods that exploit a latent factor structure, such as dynamic factor models (e.g. Stock and Watson, 2002a,b; Bai and Ng, 2008; Eickmeier and Ziegler, 2008), weighted principal components (Boivin and Ng, 2006), sparse principal components (Kristensen, 2015) or factor augmented vector autoregressions (Bernanke et al., 2005; Pesaran et al., 2011; Bai et al., 2015). 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 and Watson, 2002a, 2012), the U.K. (Artis et al., 2005) and the Euro area (Marcellino et al., 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 (Gelper and Croux, 2008; De Mol et al., 2008; Kim and Swanson, 2014; Li and Chen, 2014), reduced-rank vector autoregressions (Bernardini and Cubadda, 2015), bayesian vector autoregressions (Bańbura et al., 2010) and penalized vector autoregressions (Hsu et al., 2008; Callot and Kock, 2014; Kascha and 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 comparable or superior 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 series 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, 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).

Shrinkage estimators

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

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

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

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

Given a $\alpha \in (0, 1]$ and a large enough value for λ these estimators, from heron 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 from irrelevant variables with small but non-zero coefficients. 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) demonstrate 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 et al. (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 and 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 et al. (2007) who apply the (adaptive) lasso to models with autoregressive errors and derive estimation and selection consistency, and Yoon et al. (2013) who build on the results of Wang et al. 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, but they allow the number of relevant variables to increase. They provide non-asymptotic bounds and sufficient conditions for asymptotic consistency of the predictions, parameter estimation 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 residuals. These results hold for (conditionally) heteroskedastic processes as well, although Wagener and Dette (2013) demonstrate that in the case of classical heteroskedasticity the adaptive lasso estimates possesses a suboptimal asymptotic variance. To remedy this loss in efficiency, Wagener and Dette propose a weighted version of the adaptive lasso that obtains a lower asymptotic variance and this approach is recently extended by Ziel (2016) to also cover conditional heteroskedasticity by iteratively reweighting the adaptive lasso to exploit the variance structure. 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.

Factor models

The method of principal components has among others been popularized due to the work of Stock and Watson (2002a,b) and Bai and Ng (2006), and its use is motivated through the conceptualization of factors as unobserved 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 $\boldsymbol{X}$ with a smaller number of factors:

$$x_{it} = \lambda_i(L)' \boldsymbol{f}_t + e_{it} \tag{4}$$

for $i = 1, \ldots, N$ where x_{it} is the time series i observed at time t normalized to have unit variance and zero mean, $\boldsymbol{f}_t$ is an $(s \times 1)$ vector containing common factors and e_{it} is an idiosyncratic disturbance. Furthermore, $\lambda_i(L)$ is a lag polynomial of finite order q such that the model can be rewritten in static form as

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

where $\boldsymbol{F}_t = (\boldsymbol{f}_t', \ldots, \boldsymbol{f}_{t-q}')'$ is a vector of size r with $r \leq (q+1)s$ and $\boldsymbol{e}_t = (e_{1t}, \ldots, e_{Nt})'$. Hence, factor models estimate our generic model (1) by fitting

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

with $\boldsymbol{\beta}_F = \boldsymbol{\Lambda}'\boldsymbol{\beta}_x$ and u_{t+h} being the composite error that includes the idiosyncratic error 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 $\hat{y}_{T+h|T} = \boldsymbol{w}_T'\hat{\boldsymbol{\beta}}_w + \hat{\boldsymbol{F}}_T'\hat{\boldsymbol{\beta}}_{\hat{F}}$.

The static representation of the approximate factor model (5) allows for estimation of the factors with a wide variety of principal component type estimators. For any given k, which need not be equal to the true number of 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

$$\left(\hat{\boldsymbol{\Lambda}}^k, \hat{\boldsymbol{F}}^k\right) = \underset{\boldsymbol{\Lambda}^k, \boldsymbol{F}^k}{\arg\min} \sum_t (\boldsymbol{x_t} - \boldsymbol{\Lambda^k}\boldsymbol{F}_t^{\boldsymbol{k}})'\boldsymbol{\Omega}^{-1}(\boldsymbol{x_t} - \boldsymbol{\Lambda^k}\boldsymbol{F}_t^{\boldsymbol{k}}) \tag{6}$$

with $\boldsymbol{\Omega} = \boldsymbol{I}_N$ and subject to the normalization $\boldsymbol{\Lambda}^{k\prime}\boldsymbol{\Lambda}/N = \boldsymbol{I}_k$ and $\boldsymbol{F}^{k\prime}\boldsymbol{F}$ 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 non-sphericity of the error terms in (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 (6) with a feasible estimate $\hat{\boldsymbol{\Omega}}^{-1}$. Boivin and Ng (2006, p. 185) propose several weighting rules to obtain such feasible estimates and throughout this paper we will report results based on their weighting "rule SWa", where $\hat{\boldsymbol{\Omega}}^{-1}$ is diagonal with the i^{th} diagonal element equal to $\left(\frac{1}{NT}\sum_{t=1}^{T}\hat{\boldsymbol{e}}_t\hat{\boldsymbol{e}}_t'\right)_{ii}^{-1}$. We explore the five additional rules proposed in their original paper as well, but for the sake of brevity only report those results for which a noteworthy difference in performance is observed.

Another disadvantage of principal component analysis is that every component is a linear combination of all variables in the original dataset, thereby significantly impeding interpretability of the components. In empirical applications it is commonly observed that in every component large groups of variables carry small, non-zero loadings and consequently have a negligible effect (e.g. Stock and Watson, 2002b; Croux and Exterkate, 2011; Kristensen, 2015). In an attempt to reduce the forecasting variance by removing the noise of irrelevant variables, and improve interpretability of the components as a consequence, one can introduce sparsity in the loadings when estimating the principal components. The first to propose the method of sparse principal components (SPC) are Zou et al. (2006), who reformulate the PC estimation as a regression-type problem and

impose the ℓ_1-penalty on the regression coefficients. Shen and Huang (2008) derive a computationally more beneficial method to obtain SPC by phrasing the problem as finding a low-rank approximation to the original data matrix or residual matrix for $k > 1$. We adopt the approach of Shen and Huang and refer the reader to their original paper for details.

Finally, Bai and Ng (2008) propose a further refinement to the method of 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.[1] Having reviewed all methods considered in this paper, we now proceed to our simulation study.

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:

$$\begin{aligned} y_{t+1} &= \boldsymbol{x}_t'\boldsymbol{\beta}_x + \sqrt{\theta}\epsilon_{t+h} \\ (1-\alpha L)\epsilon_{t+1} &= v_{t+1} \end{aligned} \tag{7}$$

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

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

[1] 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. Stock and Watson, 2012; Kim and Swanson, 2014). This approach is not pursued in the current paper.

which allows for regulation of the degree of pairwise correlation 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 $\hat{y}_{T+1|T} = \boldsymbol{x}_T'\hat{\boldsymbol{\beta}}_x$. 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 to select the optimal value(s). Time series cross-validation 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, 2016). To illustrate, define 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 $\hat{\boldsymbol{\beta}}(\lambda_j)$. Following, a pseudo out-of-sample mean squared forecast error is calculated as $MSFE(\lambda_j) = \frac{1}{T-c_T}\sum_{t=c_T+1}^{T}(y_t - \boldsymbol{z}_t'\hat{\boldsymbol{\beta}}(\lambda_j))^2$. This procedure is followed for all values of the tuning parameter in the predefined grid and the final tuning parameter is chosen as

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

This method is preferred over traditional k-fold cross-validation, because the time structure of the data is kept intact which is relevant when observations are dependent over time as is obviously the case in our simulations for any $\alpha > 0$.

For the principal component type methods we use the information criteria in Bai and Ng (2002) to select the number of common components to be included in the forecasting regression, with a pre-specified maximum of 10. Although we obtain results for all six information criteria, for the sake of brevity we tabulate the results for their IC_1 and PC_1 criteria only and restrict the discussion of the results on the remaining information criteria to a qualitative nature.

We generate $J = 1,000$ one-step ahead forecasts and evaluate the forecast performance of model i with the mean squared forecast error (MSFE)

$$MSFE_i = \frac{1}{J}\sum_{j=1}^{J}(y_{j,T+1} - \hat{y}_{i,j,T+1|T})^2. \tag{8}$$

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 rel-

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	1.08	1.09	1.11	1.07	1.08	**1.01**	1.04	1.07	1.09	1.01	1.05
RMSE	2.13	2.46	2.97	2.06	2.41	1.22	1.87	2.06	2.54	1.22	2.04
consistent	0%	0%	0%	27%	14%	84%	52%	27%	11%	84%	36%
conservative	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
#variables	10.00	10.00	10.00	6.38	7.79	5.20	6.04	6.38	8.01	5.20	6.87
Panel B: $N = 50$											
RMSFE	1.98	1.80	1.87	1.19	1.20	**1.03**	1.11	1.19	1.20	1.03	1.15
RMSE	18.77	15.58	17.33	4.79	4.58	1.60	3.33	4.79	4.66	1.60	3.86
consistent	0%	0%	0%	12%	4%	64%	22%	12%	4%	64%	15%
conservative	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
#variables	50.00	50.00	50.00	8.29	15.30	5.82	12.15	8.33	15.53	5.82	16.98
Panel C: $N = 100$											
RMSFE	-	-	7.97	1.34	1.27	1.08	1.10	1.34	1.28	**1.08**	1.12
RMSE	-	-	138.02	6.53	5.85	2.57	2.98	6.57	5.89	2.63	3.23
consistent	-	-	0%	7%	2%	38%	15%	7%	2%	37%	12%
conservative	-	-	100%	100%	100%	100%	100%	100%	100%	100%	100%
#variables	-	-	100.00	9.42	19.55	6.42	10.61	9.50	19.73	6.51	11.06

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

evant variables only. As a measure of the estimation accuracy we calculate the mean squared error of the coefficient vectors as

$$MSE_i = \frac{1}{J}\sum_{j=1}^{J}\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{i,j}\right\|_2^2, \tag{9}$$

and, again, report the MSE relative to the OLS oracle procedure. Given the misspecified nature of the principal component estimators on the set of DGPs under consideration, this metric is reported for the shrinkage estimators only.

The selection performance 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 table 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 improve 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 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 and 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 expectations, given the proclaimed benefits of ℓ_2-penalization under multi-collinearity, but 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. Furthermore, Panel B clearly depicts the deterioration in selection performance after the introduction of 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

Table 2 Stationary observed variables: the effect of correlation

		ols	Ridge		Las		AdaLas		EN		AdaEN	
ρ	α		BIC	CV	BIC	CV	BIC	CV	BIC	CV	BIC	CV
Panel A: RMSFE												
0.0	0.0	1.98	1.80	1.87	1.19	1.20	**1.03**	1.11	1.19	1.20	1.03	1.15
0.6	0.0	1.96	1.57	1.61	1.16	1.19	1.07	1.13	1.16	1.20	**1.07**	1.15
0.6	0.6	1.77	1.48	1.60	1.17	1.18	1.08	1.11	1.17	1.19	**1.08**	1.13
Panel B: Consistent												
0.0	0.0	0%	0%	0%	12%	4%	64%	22%	12%	4%	64%	15%
0.6	0.0	0%	0%	0%	5%	2%	42%	16%	5%	1%	42%	10%
0.6	0.6	0%	0%	0%	4%	2%	43%	14%	4%	2%	43%	10%
Panel C: #variables												
0.0	0.0	50.00	50.00	50.00	8.29	15.30	5.82	12.15	8.33	15.53	5.82	16.98
0.6	0.0	50.00	50.00	50.00	9.18	15.11	6.26	11.23	9.19	15.89	6.27	16.12
0.6	0.6	50.00	50.00	50.00	9.31	15.48	6.29	11.75	9.31	16.12	6.29	16.34

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).

of collinearity lead to larger models with irrelevant variables being erroneously included at higher frequencies. Finally, the introduction of serial correlation has little effect on the relative forecasting or selection performance. The method by which we scale the idiosyncratic noise term controls for the increased variance induced by serial correlation and, as a result, the RMSFE obtained over k trials varies only marginally.

Finally, in table 3 we examine the predictive capabilities of factor models in the current framework. The LA(PC)-B estimator first proposes a subset of variables by applying the lasso estimator with the tuning parameter optimized by BIC to the candidate set. Alternatively, the LA(PC)-A estimator optimizes by the AIC-criterion and is accordingly expected to propose larger subsets. Unsurprisingly, on a DGP absent of common components the PC-type estimators display inferior performance compared to the shrinkage estimators (panel A). While the forecast accuracy worsens less when the variables in the dataset are correlated 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 leads to a loss of information that negatively affects the forecasting performance. In line with Bai and Ng (2002), the PC_1 criterion selects more components on average (panel B), thus resulting in better forecasting performance on DGPs absent of common components.

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 (2015):

Table 3 Stationary observed variables: factor models

		PC		WPC - SWa		SPC		LA(PC)-B		LA(PC)-A	
N	ρ	IC1	PC1	IC1	PC1	IC1	PC1	IC1	PC1	IC1	PC1
Panel A: RMSFE											
10	0.0	8.99	**1.08**	9.33	1.08	9.25	1.08	8.56	3.61	9.21	1.98
10	0.6	2.43	**1.07**	4.38	1.07	2.55	1.07	1.95	1.10	1.95	1.08
50	0.0	9.49	9.49	9.37	9.37	9.50	9.50	**7.92**	**7.92**	9.33	9.33
50	0.6	4.78	2.74	4.98	2.80	4.77	2.75	3.66	3.05	4.08	**2.67**
100	0.0	9.86	9.86	10.18	10.18	9.97	9.97	**8.41**	**8.41**	10.13	10.13
100	0.6	5.09	4.07	5.16	4.11	5.05	4.19	3.90	3.90	4.63	**3.87**
Panel B: #Variables											
10	0.0	1.00	10.00	1.00	10.00	1.00	10.00	1.00	4.73	1.00	7.00
10	0.6	1.16	10.00	1.34	10.00	1.17	10.00	1.06	4.76	1.09	6.21
50	0.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
50	0.6	1.00	8.86	1.01	8.13	1.00	8.81	1.00	1.61	1.00	4.98
100	0.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
100	0.6	1.00	5.92	1.00	5.62	1.00	5.02	1.00	1.03	1.00	4.84

Notes: see notes in 1. The metrics considered are the RMSFE (panel A) and the number of components (panel B). Within each panel the different rows correspond to different settings of dimension of the candidate set (N) and serial correlation (ρ).

$$\begin{aligned} x_{it} &= \boldsymbol{\lambda}_i' \boldsymbol{f}_t + e_{it} \\ (1-\alpha L)e_{it} &= (1+\theta^2)v_{it} + \theta v_{i+1,t} + \theta v_{i-1,t} \end{aligned} \tag{10}$$

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 error term and is generated from a standard normal distribution. We impose sparsity in the component loadings by setting a fraction τ of them equal to zero. The variable to forecast is generated as

$$y_t = \boldsymbol{f}_t' \boldsymbol{\beta}_f + \epsilon_t \tag{11}$$

where $\boldsymbol{\beta}$ 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 PC-type estimators use the extracted components as $\hat{y}_{T+1|T} = \hat{\boldsymbol{f}}_t' \hat{\boldsymbol{\beta}}_f$. Forecasting performance is measured on the basis of the MSFE relative to the factor-augmented regressions with the true number of components, 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 (2015), 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

$$R^2 = \frac{Tr\left(\boldsymbol{F}'\hat{\boldsymbol{F}}(\hat{\boldsymbol{F}}'\hat{\boldsymbol{F}})^{-1}\hat{\boldsymbol{F}}'\boldsymbol{F}\right)}{Tr\left(\boldsymbol{F}'\boldsymbol{F}\right)}, \tag{12}$$

where $\hat{\boldsymbol{F}} = (\hat{\boldsymbol{f}}_1, \ldots, \hat{\boldsymbol{f}}_T)'$. 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.

Table 4 - panel A highlights that the forecasting performance of the principal component type estimators is superior to OLS or shrinkage estimation on a DGP where the population covariance matrix of the disturbance terms is diagonal, i.e. $\alpha = 0$ and $\theta = 0$. The IC_1 does a good job in identifying the presence of a single common factor, whereas the PC_1 criterion has a tendency to overestimate the number of factors. The trace R^2s are close to unity, implying accurate recovery of a rotation of the unobserved factor by the PC-type estimators. The OLS estimator obtains the lowest forecast accuracy of all methods, while the shrinkage estimators perform better than OLS, but substantially worse than the PC-type estimators.

The result that standard principal components is optimal on a DGP with common factors absent of non-sphericity is not surprising given the asymptotic equivalence of maximum likelihood estimation and the method of principal components under the assumption of normality (e.g. Chamberlain and Rothschild, 1983). Perhaps more interesting is the finding that the shrinkage estimators seem to forecast more accurately than the traditional OLS estimator, implying some sort of favourable subset selection that reduces the forecast error. To the best of our knowledge, theoretical results for this phenomenon are not available in the current literature, although the recent contribution by De Mol et al. (2008) offers an intuitive conjecture. First of all, ridge regression can be viewed as a smoothed version of principal components and by limiting the allowed growth rate of the maximum eigenvalue of the population covariance matrix of the idiosyncratic component in (5), De Mol et al. (2008); Carrasco and Rossi (2016) are able to show prediction consistency of the ridge estimator on DGPs with an approximate factor structure. However, these results do not extend to lasso-type estimators. In this case, De Mol et al. (2008) postulate that collinearity in the candidate set favourably contributes to the forecasting performance. Under collinearity a few appropriately selected variables could capture the majority of the covariance in the data and span approximately the same space as the common factors.

The subset of the data proposed by methods employing an ℓ_1-penalty offers merely an approximation to the factor space and, since the selection performance of these estimators worsens under high degrees of collinearity, a consequence of the proposition by De Mol et al. (2008) is that forecasts from lasso-type estimators should not be expected to outperform correctly specified factor-augmented regressions. Indeed, panel A-B of table 4 show that the shrinkage estimators underperform the PC-type estimators, regardless of whether the component loadings are sparse. However, in panel C we observe that, after the introduction of non-sphericity in the factor DGP, the forecasting performance is tilted in favour of the shrinkage estimators. Under high levels of non-sphericity the PC-type estimators have difficulty in accurately estimating the unobserved factors, as indicated by the decrease in trace R^2s, thereby resulting in inferior forecasting performance. This same pattern is observed in the DGP with four factors, the results of which are displayed in table 5. Upon further analysis, the introduction of cross-sectional

Table 4 DGP with one common factor

				PC		WPC - SWa		SPC		LA(PC)-B		ols	Ridge	Las	AdaLas	EN	Adaen
	α	θ	τ	IC1	PC1	IC1	PC1	IC1	PC1	IC1	PC1		BIC	BIC	BIC	BIC	BIC
RMSFE	0	0	0	1.00	1.01	**0.98**	**0.98**	1.00	1.00	1.10	1.10	1.90	1.40	1.23	1.33	1.23	1.32
nvar	0	0	0	1.00	2.05	1.00	1.00	1.00	1.24	1.00	1.00	50.00	50.00	14.22	9.45	14.23	9.51
RSQ	0	0	0	0.96	0.96	0.97	0.97	0.96	0.96	0.96	0.96	-	-	-	-	-	-
RMSFE	0	0	0.4	1.00	1.00	**0.98**	**0.98**	0.98	0.98	1.10	1.10	1.81	1.37	1.17	1.23	1.17	1.23
nvar	0	0	0.4	1.00	1.37	1.00	1.00	1.00	1.05	1.00	1.00	50.00	50.00	11.48	8.11	11.50	8.13
RSQ	0	0	0.4	0.94	0.94	0.95	0.95	0.95	0.95	0.94	0.94	-	-	-	-	-	-
RMSFE	0.5	1	0.4	1.00	0.68	1.05	0.69	1.00	0.72	0.87	0.62	0.32	**0.28**	0.28	0.28	0.28	0.28
nvar	0.5	1	0.4	1.00	9.91	1.00	8.69	1.00	9.51	1.00	5.70	50.00	50.00	24.98	18.85	25.05	18.84
RSQ	0.5	1	0.4	0.41	0.72	0.42	0.71	0.42	0.70	0.52	0.72	-	-	-	-	-	-

Notes: Numerical entries in this table are averages obtained over 1,000 simulations relative to the PC estimator that uses a single components in the forecasting equation. The metrics considered are listed in the first column, whereas the following three columns describe the settings for 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 - SWa		SPC		LA(PC)-B		ols	Ridge	Las	AdaLas	EN	Adaen
	α	θ	τ	IC1	PC1	IC1	PC1	IC1	PC1	IC1	PC1		BIC	BIC	BIC	BIC	BIC
RMSFE	0	0	0	1.19	1.01	1.19	**0.98**	1.19	0.99	1.21	1.17	1.74	1.15	1.25	1.28	1.25	1.28
nvar	0	0	0	3.87	5.75	3.90	4.23	3.87	4.96	3.64	4.14	50.00	50.00	13.27	10.83	13.27	10.84
RSQ	0	0	0	0.94	0.96	0.94	0.97	0.94	0.96	0.80	0.85	-	-	-	-	-	-
RMSFE	0	0	0.4	2.06	1.03	1.69	**0.90**	2.03	1.01	1.24	1.11	1.76	1.23	1.18	1.26	1.18	1.26
nvar	0	0	0.4	3.27	5.78	3.38	4.00	3.28	4.78	3.14	4.01	50.00	50.00	15.00	11.76	15.03	11.75
RSQ	0	0	0.4	0.80	0.94	0.83	0.95	0.80	0.94	0.71	0.84	-	-	-	-	-	-
RMSFE	0.5	1	0.4	1.46	0.79	1.46	0.82	1.46	0.86	1.43	0.84	0.35	**0.32**	0.33	0.34	0.33	0.34
nvar	0.5	1	0.4	1.01	9.80	1.01	8.22	1.01	9.41	1.01	7.00	50.00	50.00	31.34	24.05	31.35	24.09
RSQ	0.5	1	0.4	0.20	0.69	0.19	0.66	0.20	0.67	0.18	0.58	-	-	-	-	-	-

Notes: Numerical entries in this table are averages obtained over 1,000 simulations relative to the PC estimator that uses four components in the forecasting equation. The metrics considered are listed in the first column, whereas the following three columns describe the settings for the degree of serial correlation (α), cross-sectional correlation (θ) and sparsity in the loadings (τ).

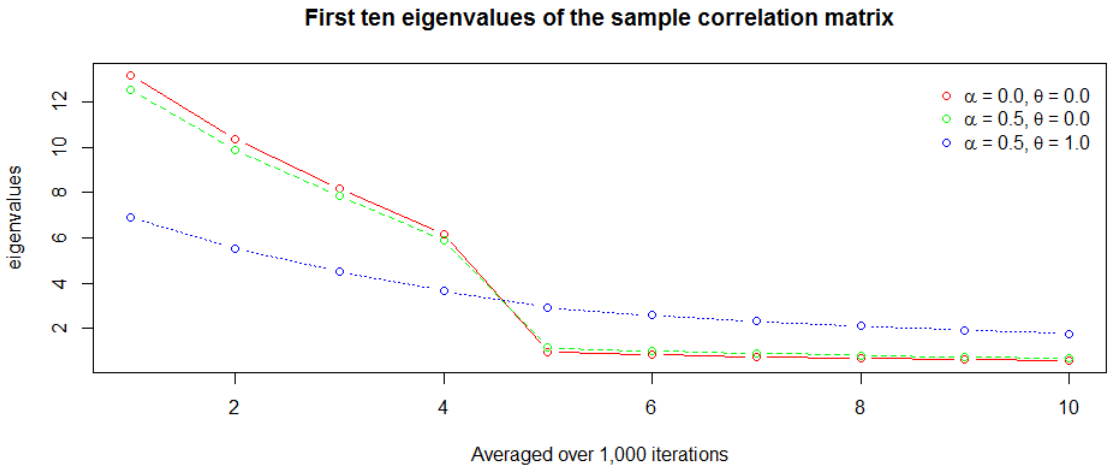

Figure 1: Visualization of the explanatory power of the first ten common components.

correlation in the error term in (10) 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 figure 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 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.

The unreported MSFE of the shrinkage estimators remains relatively stable under varying levels of non-sphericity in (10), resulting in a strong comparative advantage over the PC-type estimators. Most remarkable is the similarity in forecasting performance between Ridge and the lasso-type methods, despite the latter removing approximately half of all variables from the candidate set on average. In the conducted simulations Ridge regression tends to give equal weight to each variable in the candidate set on average, in line with the random generation of the component loadings that do not consistently overweight the effect of specific factors on a specific subset of the data. The lasso-type regressions, however, show a more variable spread in the weights given to each variables, resulting from the positive probability mass these methods put at zero coefficient values. Indeed, we observe that when Ridge estimates a coefficient that is small in magnitude, the corresponding lasso-type estimate tends to be exactly zero. The latter finding is additionally reflected in the relatively low correlations between the coefficient estimates of Ridge and the lasso-type methods, which for most coefficients lies in a neighbourhood of 0.5.

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 is has recently been explored by Wilms and Croux (2016), Liao and Phillips (2015) and Liang and Schienle (2015) who all explore the use of penalized regression in automated vector error correction model estimation. These novel and insightful contributions, however, require the implementation of a non-standard and fairly technical model. 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:

$$\begin{aligned} \Delta y_t &= \alpha \left(y_{t-1} - \sum_{i=1}^{3} \beta_i x_{i,t-1} \right) + \epsilon_{j,t} \\ x_{i,t} &= x_{i,t-1} + \epsilon_{j+1,t} \qquad\qquad i = 1,2,3,\ j = 1,2,3 \end{aligned} \tag{13}$$

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 cointegrated variables $x_{i,t}$ for $i = 1, \ldots, 3$, we fill the candidate set $\boldsymbol{X}$ with a number of irrelevant variables. 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 (7) 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 1,000 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 principal component estimators is excluded from this section on the grounds that every extracted component will contain a linear combination of all variables and, hence, will be integrated of order one. 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 et al. (2014), or estimation of the factors in a VECM framework in the spirit of Barigozzi et al. (2016a,b). A preliminary analysis confirms that the regular factor methods considered in this paper all display sub-par performance and are therefore omitted from the current analysis, while we postpone a comparative analysis including factor-augmented error correction model to future research. 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 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 exactly identify the sparse cointegrating relationship. 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 reveal 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

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
Lasso	1.17	1.51	1.74	1.17	1.58	1.82
Ada-Lasso	**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
Ada-EN	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
Lasso	4.22	8.21	10.64	5.31	18.88	26.90
Ada-Lasso	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
Ada-EN	2.16	3.24	8.08	2.52	16.46	25.14
Panel C: Consistent						
Lasso	29.9%	20.1%	18.2%	9.8%	0.2%	0.0%
Ada-Lasso	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%
Ada-EN	81.2%	62.2%	33.5%	63.6%	4.1%	0.2%
Panel D: Conservative						
Lasso	99.5%	93.1%	88.5%	99.6%	82.5%	64.1%
Ada-Lasso	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%
Ada-EN	99.8%	99.6%	91.6%	99.9%	79.3%	58.2%
Panel E: #Variables						
Lasso	4.53	6.29	6.65	5.35	9.97	12.17
Ada-Lasso	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
Ada-EN	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.

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	6.58	15.88	1.46	7.42	22.16	3.13
MSFE	140.30	65.81	22.01	167.07	98.33	45.37
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	0.51	0.55	0.42	0.95	0.99	0.80
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).

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 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 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 slow, e.g. $\alpha = -0.1$, the long run dependence may be so high 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 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 and ADF-test in panel B all obtain substantially higher RMSFEs. Furthermore, for all adjustment rates these models almost never contain all relevant variables. This provides a strong 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 seem to do better when there exists strong forms of long run dependence. 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.

In conclusion, the use of lasso-type estimators on a high-dimensional non-stationary dataset containing cointegrated variables provides forecast gains over traditional approaches such as the use of OLS to estimate the cointegrating vector on the entire

dataset or pre-processing the data to stationarity. 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 the necessity to validate 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 based 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. [2] In congruence with our previous argument on the unknown consequences brought forward by the necessity of validation the presence of cointegration in empirical datasets, we 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. Stock and Watson, 2002b; Ludvigson and Ng, 2009; Kristensen, 2015).

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

$$y_{t+h}^h = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i+1} + \epsilon_{t+h}, \tag{14}$$

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 $p \in \{0, \dots, 6\}$, is determined by the BIC criterion, as is the case for all following methods. The penalized regressions obtain the forecasts by fitting

$$y_{t+h}^h = \alpha + \boldsymbol{x}_t' \boldsymbol{\beta}_x + \sum_{i=1}^{p} \beta_i y_{t-i+1} + \epsilon_{t+h}, \tag{15}$$

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

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, the principal component regressions fit

$$y_{t+h}^h = \alpha + \hat{\boldsymbol{F}}_t' \boldsymbol{\beta}_F + \sum_{i=1}^{p} \beta_i y_{t-i+1} + \epsilon_{t+h}, \tag{16}$$

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).

We simulate real-time forecasting by calculating pseudo out-of-sample forecasts at horizons $h = 1, 6$ and 12. An initial in-sample period covering 12 years of monthly observation is used to estimate the models by which to obtain the first out-of-sample prediction. For each new prediction, we either keep the length of the in-sample period fixed, resulting in a rolling window approach, or we let the in-sample period expand with each prediction under a recursive scheme. Regardless of the scheme employed, 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.

Results for the rolling window forecasts are reported in table 8 and for the recursive window forecasts in table 9. The first row in each panel depicts the actual MSFEs obtained by the AR benchmark, whereas the remaining rows are MSFEs relative to the benchmark. The tuning parameters for each method are selected by a variety of methods: BIC, AIC and time series CV for penalized regression and all six Bai and Ng (2002) information criteria for the PC-type methods and hybrid methods. We report the lowest MSFE among all tuning parameter selection approaches and additionally provide an overview of the performance of each approach in table 10. Finally, table 11 summarizes the overall forecasting results by tabulating the "winning" methods per forecast horizon.

A notable pattern in table 8 and 9 seems to be that the method of principal components, either in combination with shrinkage or without, tends to display superior performance on the real series, whereas pure penalized regression tends to perform better on the price indices. Furthermore, with the exception of the recursive PPI 12-step-ahead forecasts, the AR benchmark is never optimal. In the majority of cases, the forecast errors are lower for the recursive estimates, thus demonstrating the added value of utilizing more historical data despite the increased risk of estimating models on data containing structural breaks. No clear differences are observed in the relative performance between methods based on the choice of rolling or recursive estimation. To provide insights into the performance of the tuning methods adopted, we report an overview of the fraction of cases in which each method obtains the lowest MSFE in table 10. For each method, a total of 64 series-horizon combinations are forecast. Focussing on the shrinkage methods in panel A, we observe that the size of the ℓ_2-penalty in Ridge regression is best optimized by time series cross-validation, which is unsurprising given that Ridge does not perform

Table 8 Empirical application: rolling window

Method	RPI	IP	RMTS	EMP	PPI	CPIA	CPIL	PCEPI
Panel A: h = 1								
AR	48.76	66.05	130.73	3.00	54.74	14.37	17.51	9.23
Ridge	1.00	0.93	1.02	0.85	0.96	0.76	0.77	0.74
Lasso	0.99	0.89	0.95	0.85	0.94	0.71	0.72	0.67
AdaLasso	0.97	0.87	0.93	0.87	0.97	0.70	**0.70**	**0.66**
EN	0.99	0.89	0.98	0.87	0.94	0.71	0.72	0.67
AdaEN	0.97	**0.86**	0.93	0.86	0.97	**0.69**	0.70	0.66
PC	0.97	0.88	**0.86**	0.85	0.91	0.75	0.76	0.73
WPC-SWa	0.98	0.90	0.86	0.86	0.91	0.75	0.77	0.73
WPC-SWb	0.97	0.88	0.86	0.82	0.90	0.70	0.71	0.67
WPC-1	0.97	0.89	0.87	**0.80**	0.92	0.76	0.78	0.73
WPC-2	**0.97**	0.87	0.86	0.83	0.93	0.79	0.81	0.81
SPC	0.97	0.89	0.86	0.85	0.92	0.75	0.76	0.74
LA(PC)-B	1.01	0.89	0.94	0.86	**0.90**	0.72	0.71	0.71
LA(PC)-A	1.00	0.89	0.88	0.84	0.92	0.75	0.76	0.72
Panel B: h = 6								
AR	6.90	30.88	31.16	2.28	1.46	0.41	0.51	0.25
Ridge	0.78	0.81	0.74	0.83	1.02	1.05	1.09	0.98
Lasso	0.91	0.90	0.80	0.85	1.05	0.90	**0.87**	**0.93**
AdaLasso	0.94	0.90	0.82	0.82	1.08	0.91	0.90	0.99
EN	0.87	0.90	0.79	0.83	1.05	**0.90**	0.88	0.93
AdaEN	0.90	0.86	0.82	0.83	1.08	0.91	0.90	0.98
PC	0.78	0.78	0.68	0.80	1.06	1.02	1.01	1.02
WPC-SWa	0.82	0.91	0.69	0.84	1.03	0.97	0.97	0.98
WPC-SWb	0.79	0.84	0.68	0.82	1.03	0.96	0.96	0.96
WPC-1	**0.77**	0.82	0.71	0.81	1.07	1.00	1.01	0.98
WPC-2	0.78	0.85	0.73	0.85	1.03	1.02	1.02	1.00
SPC	0.78	0.79	0.68	0.81	1.06	1.03	1.01	1.02
LA(PC)-B	0.88	0.81	0.73	0.84	**0.99**	0.92	0.91	0.95
LA(PC)-A	0.86	**0.78**	**0.67**	**0.80**	1.07	1.01	1.01	1.00
Panel C: h = 12								
AR	4.53	23.58	20.92	2.60	0.34	0.10	0.12	0.06
Ridge	**0.75**	0.81	0.82	0.85	1.11	1.02	1.04	1.00
Lasso	0.90	0.88	0.91	0.98	**0.98**	0.91	**0.93**	0.86
AdaLasso	0.95	0.90	0.93	1.07	1.02	0.90	0.94	0.89
EN	0.90	0.86	0.91	0.97	0.99	0.91	0.93	**0.85**
AdaEN	0.96	0.89	0.94	1.07	1.02	**0.90**	0.94	0.89
PC	0.80	0.73	0.70	0.73	1.02	1.02	1.03	1.00
WPC-SWa	0.86	0.95	0.78	0.79	1.03	0.98	1.01	1.00
WPC-SWb	0.78	0.83	0.71	0.76	1.00	0.97	0.98	0.95
WPC-1	0.82	0.83	0.75	0.78	1.04	1.01	1.05	0.96
WPC-2	0.80	0.82	0.76	0.79	1.03	1.01	0.99	0.99
SPC	0.82	**0.71**	0.71	**0.70**	1.01	0.98	1.02	0.98
LA(PC)-B	0.82	0.96	0.96	0.97	1.03	0.92	0.98	0.91
LA(PC)-A	0.80	0.73	**0.69**	0.74	1.03	1.02	1.03	1.03

Notes: Numerical entries in this table are pseudo out-of-sample mean squared forecast errors relative to a univariate AR benchmark. Forecasts are based on a rolling window etimation scheme.

Table 9 Empirical application: recursive window

Method	RPI	IP	RMTS	EMP	PPI	CPIA	CPIL	PCEPI
Panel A: h = 1								
AR	48.61	64.78	130.11	2.88	53.38	16.10	18.91	10.65
Ridge	0.95	0.91	0.90	1.00	0.92	0.65	0.67	0.87
Lasso	0.98	0.88	0.88	0.91	0.90	0.56	0.61	0.53
AdaLasso	0.98	**0.83**	0.90	0.90	0.91	0.56	0.62	0.55
EN	0.95	0.88	**0.83**	0.91	0.89	0.56	0.63	0.53
AdaEN	0.95	0.84	0.95	0.96	0.91	**0.56**	**0.61**	**0.53**
PC	**0.92**	0.87	0.89	**0.86**	0.87	0.64	0.67	0.65
WPC-SWa	0.93	0.88	0.93	0.91	0.90	0.73	0.77	0.75
WPC-SWb	0.93	0.88	0.90	0.87	0.89	0.69	0.73	0.69
WPC-1	0.93	0.89	0.90	0.87	0.88	0.65	0.70	0.65
WPC-2	0.94	0.88	0.96	0.89	0.93	0.73	0.76	0.74
SPC	0.93	0.88	0.89	0.86	0.87	0.64	0.67	0.65
LA(PC) - B	0.97	0.85	0.93	0.96	0.86	0.61	0.62	0.56
LA(PC) - A	0.96	0.87	0.90	0.89	**0.84**	0.60	0.69	0.61
Panel B: h = 6								
AR	6.86	30.10	28.08	2.27	1.43	0.45	0.55	0.29
Ridge	0.86	0.81	0.75	0.93	**0.95**	**0.78**	0.83	**0.76**
Lasso	0.95	0.80	0.83	0.88	1.06	0.81	**0.81**	0.79
AdaLasso	0.88	0.81	0.85	0.89	1.08	0.85	0.86	0.86
EN	0.88	0.78	0.78	0.88	1.06	0.80	0.81	0.79
AdaEN	**0.86**	0.78	0.81	0.89	1.08	0.83	0.84	0.86
PC	0.87	0.76	0.71	0.79	0.99	0.86	0.88	0.85
WPC-SWa	0.89	0.85	0.80	0.90	1.01	0.89	0.90	0.89
WPC-SWb	0.91	0.83	0.75	0.84	0.98	0.84	0.86	0.83
WPC-1	0.86	**0.73**	0.74	0.80	0.99	0.86	0.90	0.85
WPC-2	0.94	0.91	0.80	0.91	0.99	0.89	0.89	0.86
SPC	0.87	0.76	**0.70**	**0.79**	0.98	0.85	0.88	0.84
LA(PC) - B	0.95	0.79	0.79	0.81	1.02	0.86	0.87	0.84
LA(PC) - A	0.93	0.75	0.79	0.80	1.02	0.86	0.93	0.84
Panel C: h = 12								
AR	4.48	23.10	19.51	2.72	**0.33**	0.11	0.13	0.07
Ridge	1.00	**0.75**	0.77	0.89	1.48	0.95	0.96	0.90
Lasso	0.91	0.78	0.76	0.87	1.04	0.88	0.90	**0.78**
AdaLasso	0.94	0.82	0.85	0.97	1.21	0.90	0.95	0.83
EN	0.88	0.76	0.77	0.87	1.04	**0.88**	**0.90**	0.78
AdaEN	0.96	0.79	0.79	0.94	1.22	0.90	0.96	0.83
PC	0.89	0.71	0.71	0.79	1.04	0.90	0.93	0.89
WPC-SWa	0.95	0.82	0.76	0.88	1.08	0.94	0.95	0.93
WPC-SWb	0.93	0.75	**0.70**	0.84	1.02	0.90	0.92	0.88
WPC-1	**0.88**	0.71	0.72	0.80	1.04	0.90	0.93	0.88
WPC-2	0.98	0.89	0.76	0.91	1.02	0.91	0.92	0.89
SPC	0.89	**0.71**	0.70	0.80	1.04	0.90	0.92	0.88
LA(PC) - B	0.89	0.77	0.78	0.87	1.12	0.96	0.96	0.87
LA(PC) - A	0.89	0.72	0.74	**0.78**	1.08	0.91	0.93	0.86

Notes: see notes in 8. Forecasts are based on a recursive window etimation scheme.

Table 10 Empirical application: tuning methods

Tuning methods							
Panel A: Shrinkage estimators							
Tuning:	BIC	AIC	CV				
Ridge	4%	4%	**91%**				
Lasso	**73%**	0%	27%				
AdaLasso	**55%**	0%	45%				
EN	**67%**	0%	33%				
AdaEN	**59%**	0%	41%				
Panel B: PC/Hybrid estimators							
Tuning:	Fixed (5)	IC1	IC2	IC3	PC1	PC2	PC3
PC	23%	11%	**33%**	8%	13%	2%	11%
WPC-SWa	4%	24%	**46%**	4%	10%	4%	6%
WPC-SWb	17%	20%	**42%**	3%	2%	0%	16%
WPC-1	17%	19%	**23%**	19%	2%	3%	17%
WPC-2	6%	30%	**45%**	5%	6%	3%	5%
SPC	22%	17%	**23%**	11%	11%	5%	11%
LA(PC)-B	**27%**	20%	19%	13%	5%	5%	13%
LA(PC)-A	16%	9%	**27%**	14%	5%	6%	23%

Notes: Numerical entries in this table depict the relative frequencies with which a tuning method delivered the best forecast performance in the empirical applications listed in table 8 and 9. The estimation methods are listed in the first column, whereas the respective tuning methods are listed in the first row of each panel.

subset selection and the information criteria are unstable in high-dimensional settings. The methods incorporating an ℓ_1-penalty on the other hand do impose sparsity and in this case the BIC seems to be the preferred tuning method, closely followed by time series cross-validation. For the PC-type estimators and hybrid estimators in panel B we compare the performance of the six Bai and Ng criteria an ad-hoc method of choosing five fixed factors. Surprisingly, it seems that the $IC2$-criterion obtains the lowest MSFE most frequently for all methods, with the exception of the hybrid method $LA(PC) - B$.

Finally, the overview of the winning methods in table 11 clearly highlights the strong performance of shrinkage estimators. In over half of all cases considered, a shrinkage estimator obtained the overall lowest MSFE, whereas in an additional twentyseven percent the combination of shrinkage and PC-type estimation delivered the best performance. These results demonstrate that the dominance of factor models in economic forecasting may not always be justified, with a clear viable alternative being the usage of shrinkage estimators or, in particular, lasso-type estimators.

5 Conclusion

In this paper we examine the forecasting performance of factor models, models involving shrinkage and hybrid models. Comprehensive simulations based on a wide variety of

Table 11 Empirical application: summary

	Rolling			Recursive			Total	
horizon	h=1	h=6	h=12	h=1	h=6	h=12	#	%
AR	0	0	0	0	0	1	1	2%
Shrinkage	4	3	5	5	5	3	25	52%
Factor	3	1	0	2	1	2	9	19%
Hybrid	1	4	3	1	2	2	13	27%

Notes: A summary table depicting the number of times a category of estimators delived the best forecast performance. Results are tabulated for different estimation schemes (rolling vs. recursive) and forecast horizons (h).

data generating processes indicate that lasso-type estimators are relatively robust against alternative DGP specifications; they naturally perform well on stationary variables with a sparse coefficient vector, but also show strong forecasting performance on data driven by approximate factor structures even the latter contains a high degree of non-sphericity in the residuals. 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 simulation-based findings, with penalized regression obtaining the lowest mean squared forecast error in 52% of all series-horizons forecast and hybrid methods outperforming in 27%. We take this as further evidence that the assumption of a common factors being persistent in macroeconomic data may not always be valid or, at a minimum, may not always be relevant for forecasting purposes given the flexibility with which lasso-type estimators can handle this type of data.

References

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

Bai, J., Li, K., and Lu, L. (2015). Estimation and inference of FAVAR models. *Journal of Business & Economic Statistics*, forthcoming.

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

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74:1133–1150.

Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146:304–217.

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

Banerjee, A. and Marcellino (2009). Factor-augmented error correction models. In Castle, J. L. and Shephard, N., editors, *The Methodology and Practice of Econometrics - A Festschrift for David Hendry*, pages 589–612. Oxford University Press, Oxford.

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

Barigozzi, M., Lippi, M., and Luciani, M. (2016a). Dynamic factor models, cointegration, and error correction mechanisms. Working Paper.

Barigozzi, M., Lippi, M., and Luciani, M. (2016b). Non-stationary dynamic factor models for large datasets. Working Paper.

Bernanke, B., Boivin, J., and Eliasz, P. (2005). Factor augmented vector autoregressions (FVARs) and the analysis of monetary policy. *Quarterly Journal of Economics*, 120:387–422.

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

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

Callot, L. A. and 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*, pages 238–268.

Carrasco, M. and Rossi, B. (2016). In-sample inference and forecasting in misspecified factor models. *Journal of Business & Economic Statistics*, (forthcoming):1–72.

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

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

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

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

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

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

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

Hyndman, R. J. (2016). *forecast: Forecasting functions for time series and linear models.* R package version 7.2.

Kascha, C. and Trenkler, C. (2015). Forecasting VARs, model selection and shrinkage. Working paper 15-07, University of Mannheim / Department of Economics.

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

Knight, K. and 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. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186:325–344.

Kristensen, J. T. (2015). Diffusion indexes with sparse loadings. *Journal of Business & Economic Statistics*, forthcoming.

Li, J. and 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. and Schienle, M. (2015). Determination of vector error correction models in higher dimensions. working paper, Leibniz Universität Hannover.

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

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

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

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

Medeiros, M. C. and 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. and Rinaldo, A. (2011). Autoregressive process modeling via the lasso procedure. *Journal of Multivariate Analysis*, 102:529–549.

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

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

Song, S. and Bickel, P. J. (2011). Large vector auto regressions. *arXiv preprint arXiv:1106.3915*.

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

Stock, J. H. and Watson, M. W. (2006). Forecasting with many predictors. *Handbook of Economic Forecasting*, 1:515–554.

Stock, J. H. and 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*, 58:267–288.

Wagener, J. and Dette, H. (2013). The adaptive lasso in high-dimensional sparse heteroscedastic models. *Mathematical Methods of Statistics*, 22:137–154.

Wang, H., Li, G., and Tsai, C. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of The Royal Statistical Society B*, 69:63–78.

Wilms, I. and Croux, C. (2016). Forecasting using sparse cointegration. *International Journal of Forecasting*, 32:1256–1267.

Yoon, Y. J., Park, C., and Lee, T. (2013). Penalized regression models with autoregressive error terms. *Journal of Statistical Computation and Simulation*, 83:1756–1772.

Zhao, P. and 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 and 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., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15:265–286.