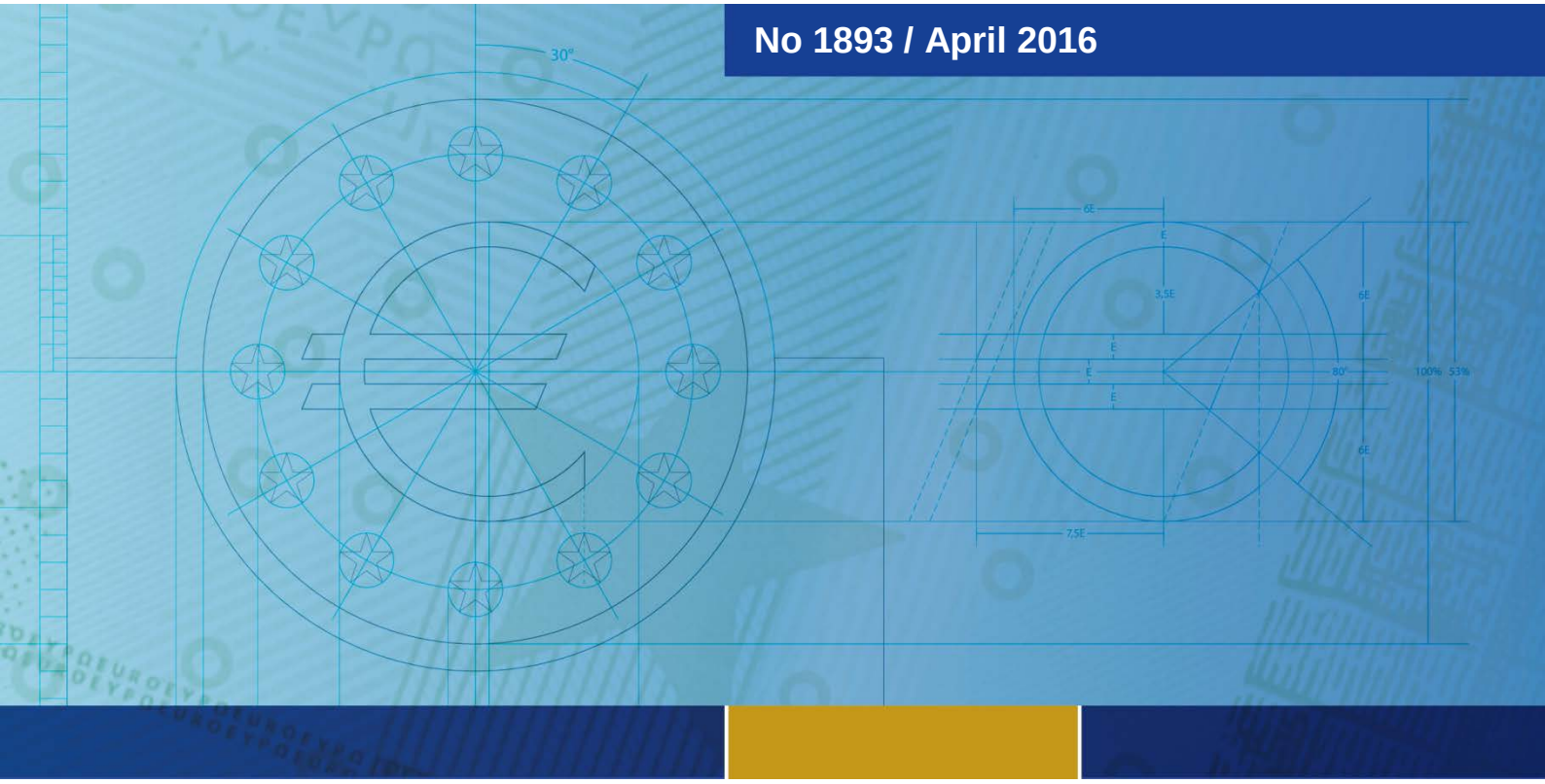
EUROPEAN CENTRAL BANK

EUROSYSTEM

Working Paper Series

Gerhard Rünstler

On the design of data sets for forecasting with dynamic factor models



No 1893 / April 2016

Abstract

Forecasts from dynamic factor models potentially benefit from refining the data set by eliminating uninformative series. The paper proposes to use prediction weights as provided by the factor model itself for this purpose. Monte Carlo simulations and an empirical application to short-term forecasts of euro area, German, and French GDP growth from unbalanced monthly data suggest that both prediction weights and Least Angle Regressions result in improved nowcasts. Overall, prediction weights provide yet more robust results.
Keywords: Dynamic factor models, forecasting, variable selection, LARS
JEL classification: E37, C53, C51

Non-technical summary

Dynamic factor models have emerged as a widely used tool for obtaining short-term forecasts of economic activity and inflation. These models are usually applied to large data sets that consist of a wide range of different series, as suggested by standard considerations from statistical theory.

Recentl studies have argued that forecasts may be improved by focusing on a limited set of highly informative series. Methods to select appropriate informative data sets are based on so-called stepwise regression, which builds an informative data set by an iterative procedure. Starting with the series with the highest information content, at each step another series (the one with the highest additional information gain in forecasting) is added to the data set. It is well-known that this procedure becomes highly inefficient, once the number of series increases. To enhance robustness, constrained versions have been proposed, the most prominent among them Least Angle Regressions (LARS).

Another way to increase robustness is to use the factor model itself for estimating the information content of individual series from their weights in the factor model prediction. In this paper, I provide two pieces of evidence, which suggest that factor model prediction weights are a useful alternative to LARS. First, a simulation exercise confirms that both methods are suitable for selecting data sets that result in more efficient predictions. However, factor model prediction weights are more successful than LARS in identifying the appropriate series and deliver more efficient predictions. Moreover LARS shows some tendency of over-fitting, as predictions suggest gains that only partly carry over to other samples.

Second, I apply both methods to the forecasting of quarterly GDP growth from large unbalanced monthly data sets from a dynamic factor model. I use monthly data sets of about 70 series for the euro area, Germany, and France over the period of 1991 to 2007. I find that variable selections from either method improve forecasts of the euro area and German GDP with small data sets of about 10 to 30 series, but not so for France. Overall, again, prediction weights provide more robust variable selections and give rise to smaller prediction errors than LARS.

1 Introduction

Dynamic factor models have emerged as a widely used tool for obtaining nowcasts and short-term forecasts of economic activity and inflation (e.g. Stock and Watson, 2002a; Gianonne et al., 2008). From asymptotic considerations, these models are usually applied to large data sets that consist of a wide range of different series. It has been questioned though that increasing the sheer number of series in the data set would necessarily improve forecast performance. Boivin and Ng (2006) have identified conditions, under which enlarging the data set may actually worsen the precision of factor estimates. Bai and Ng (2008) have proposed Least Angle Regressions (LARS) and related methods to identify efficient sets of predictors in dynamic factor models. These two studies, along with Schumacher (2010), Caggiano et al. (2011), Alvarez et al. (2012), and Bessec (2013) also present empirical applications that demonstrate gains from using smaller data sets in predictions from dynamic factor models.

In this paper, I propose the use of prediction weights that are obtained from the factor model itself as an alternative method for selecting an efficient set of predictors. As with any linear model, the factor model prediction for a certain target variable can be written as a weighted linear combination of current and lagged values of the predictors. I investigate, whether forecast efficiency can be improved by retaining only predictors with high weights.

Basically, the method parallels stepwise regression, but with the difference that a factor structure is imposed on the data. In its forward selection variant, stepwise regression builds up a set of predictors for a certain target variable by an iterative procedure. At each step, it adds the series with the highest marginal predictive gain to the set of series from the previous step. It is well-known that this procedure becomes highly inefficient, once the number of series increases. To overcome the dimensionality problem, constrained versions have been proposed, among them LARS and LASSO (Efron et al., 2004), the latter being used by Bai and Ng (2008) to select predictors in factor model forecasts of inflation. Another way to deal with high dimensionality is to use the factor model itself for estimating the marginal predictive gains of individual series. This amounts to calculating their weights in the factor model prediction.

I provide two pieces of evidence, which suggest that factor model prediction weights are a useful alternative to LARS. First, a Monte-Carlo simulation exercise confirms that both methods are suitable for selecting data sets that result in more efficient predictions. However, factor model prediction weights are more successful than LARS in identifying the appropriate series. Consequently, they also tend to deliver better out-of-sample predictions. LARS, in turn, shows some tendency of overfitting, as pre-sample predictions suggest gains that only partly carry over to the out-of-sample case.

Second, I apply both methods to the now- and forecasting of quarterly GDP growth from large unbalanced monthly data sets. I use the dynamic factor model by Doz et al. (2011), which employs a state-space framework and therefore copes with unbalanced data and mixed frequencies in an efficient way. It has been shown to perform well under these conditions (Giannone et al., 2008; Rünstler et al., 2009; Angelini et al., 2011). As pointed out by Bańbura and Rünstler (2011), prediction weights of individual series can be obtained from an extension of the Kalman filter. LARS is less suited for dealing with unbalancedness and must be applied to quarterly aggregates of monthly data.

I use monthly data sets for the euro area, Germany, and France over the period of 1991 to 2014. Each data set contains about 70 series. I obtain variable selections from a pre-sample and evaluate their performance from a pseudo real-time forecast exercise. I find that variable selections of 10 to 30 series from either method improve nowcasts of euro area GDP. Results for Germany and France are more mixed. Selections from factor model prediction weights provide moderate but consistent gains, while pre-sample LARS selections sometimes suggest gains that revert into losses in out-of-sample predictions. Overall, for nowcasts factor model predicition weights tend to provide more robust variable selections than LARS. Gains for next-quarter forecasts are generally very small.

The paper is organised as follows. Section 2 reviews the basic concepts. Section 3 discusses variable selection in the context of the dynamic factor model by Doz et al. (2011) with unbalanced and mixed-frequency data. Section 4 conducts the Monte Carlo study to investigate the gains from using prediction weights and LARS. Section 5 presents the empirical application. Section 6 concludes the paper.

2 Variable selection in factor models

Consider the dynamic factor model

$$x_t = \Lambda f_t + \xi_t, \qquad \xi_t \sim \mathbb{N}(0, \Sigma_\xi). \tag{1}$$

The model relates the $n \times 1$ vector of series $x_t = (x_{1t}, \ldots, x_{nt})'$ to $r \times 1$ vector of common factors $f_t = (f_{1t}, \ldots, f_{rt})'$ from matrix Λ of factor loadings and to $n \times 1$ vector of idiosyncratic components $\xi_t = (\xi_{1t}, \ldots, \xi_{nt})'$ with covariance matrix Σ_ξ. It holds $r << n$. Common factors f_t and idiosyncratic components ξ_t are assumed to follow certain stochastic processes, which will be specified below.

The purpose of the model is to estimate (and possibly predict) f_t from data x_t, $t = 1, \ldots, T$, and subsequently to predict a scalar target series y_t from the equation

$$y_t = \beta' f_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathbb{N}(0, \sigma_\varepsilon^2), \tag{2}$$

with $r \times 1$ vector $\beta = (\beta_1, \ldots, \beta_r)'$. Residual ε_t is assume to be identically independently distributed and to be independent of ξ_t.

As $[n; T] \to \infty$, the factor space of dimension r can be consistently estimated by principal components under various conditions, which include (i) appropriate assumptions on the stationarity and weak time dependence of f_t and ξ_t; (ii) a rank condition on Λ precluding non-trivial factor loadings; (iii) sufficiently weak cross-sectional dependence between f_t and ξ_t; and (iv) sufficiently weak cross-sectional dependence among the elements of vector ξ_t (Stock and Watson, 2002b; Bai and Ng, 2002).

Specifically, under condition (iv), the non-diagonal elements of Σ_ξ should turn sufficiently small as n tends to infinity (e.g. Bai and Ng, 2002). Boivin and Ng (2006) argue that this is likely to be violated in macro-economic data sets, as some correlation among idiosyncratic components would remain. They further show that, in finite samples, forecast precision would not necessarily increase with the number of series if Σ_ξ is non-diagonal. Under certain circumstances, e.g. with

heteroscedasticity in idiosyncratic components, or in case that some elements of f_t are irrelevant for predicting y_t, predictions may therefore be improved from removing uninformative series.

Boivin and Ng (2006) and Caggiano et al. (2011) present empirical applications, where forecasts are improved by simple removing the series with the highest cross correlations in idiosyncratic components. In search for more sophisticated selection criteria, Bai and Ng (2008) proposed Least Angle Regressions (LARS) and several variants (LASSO and elastic net algorithms) to select series in factor model forecasts for U.S. inflation. They report considerable gains in precision over a range of specifications. Schumacher (2010) and Bessec (2013) confirm these findings for German and French GDP, respectively. With the exception of Bessec (2013), these studies inspect in-sample forecasts, i.e. perform variable selection within the forecast evaluation sample. The studies use diffusion indices (Stock and Watson, 2002a).[1]

LARS is a constrained variant of stepwise forward selection to predict y_t from equation

$$y_t = \beta' x_t^{L,s} + \varepsilon_t^{L,s}, \qquad \varepsilon_t^{L,s} \sim \mathbb{N}(0, \sigma_{\varepsilon,s}^2), \tag{3}$$

where $x_t^{L,s}$ denotes a certain subset of series x_t of size s. Starting with empty set $x_t^{L,0}$, at each step s, one series is added to $x_t^{L,s-1}$ in order to obtain $x_t^{L,s}$. As with the standard approach, this is the series with the highest marginal predictive gain on top of predictions based on $x_t^{L,s-1}$, i.e. the highest correlation with residual $\varepsilon_t^{L,s-1}$. To increase the robustness of forward selection, LARS adjusts coefficients β in equation (3) after each step. This is done by increasing the coefficients in their joint least squares direction until another predictor (not yet contained in $x_t^{L,s}$) displays as much correlation with the residual as the series contained in $x_t^{L,s}$. The process stops at $k = \min(T, n-1)$, and results in a set of selections $\mathcal{L} = \left\{ x_t^{L,s} \right\}_{s=1}^{k}$.

The purpose of coefficient shrinkage in LARS is to overcome the dimensionality problem that emerges with a high number of predictors and results in highly inefficient selections. Another way to deal with high dimensionality is using the factor model itself for approximating the marginal

[1] Similarly, Barhoumi et al. (2010) and Alvarez et al. (2012) find that forecasts from small data sets that consist only of aggregate indicators outperform those from larger data sets with a high number of sectoral indicators. Taking a different perspective, Alvarez et al. (2012) and Poncela and Ruiz (2015) show that similar issues arise with the precision of factor estimates.

predictive gains of the individual series. The latter can be obtained from the weights of the individual series in the factor model predictions.

The principle is easily illustrated for a static factor model, with factors being estimated by principal components. Series x_{it} are assumed to be standardized to mean zero and variance one. Consider variable selection $x_t^{w,s}$ and let $(1/T)\sum_{t=1}^{T} x_t^{w,s}\left(x_t^{w,s}\right)' = V_s D_s V_s'$ be the eigenvalue decomposition of its empirical covariance matrix with eigenvectors V_s. Given the number of factors r, it holds $\widehat{f}_t^{(s)} = V_{s,r}' x_t^{w,s}$, where $V_{s,r}$ denotes the matrix containing the first r columns of V_s.

The prediction of y_t is then found with

$$y_{t|t}^{w,s} = \widehat{\beta}_s' V_{s,r}' x_t^{w,s} = \left(\omega_0^s\right)' x_t^{w,s}, \tag{4}$$

where $\widehat{\beta}$ is estimated from a regression of y_t on $\widehat{f}_t^{(s)}$ as from equation (2) and ω_0^s is $s \times 1$ the vector of prediction weights. These weights represent the marginal predictive gains of the elements of $x_t^{w,s}$ from projecting of y_t on f_t.

For a factor model, stepwise backward elimination seems a natural approach. Starting with the entire set of series $x_t^{w,n} = x_t$ at each step $s = n, \ldots, 1$, the factor model is re-estimated based on series $x_t^{w,s}$ and the series with the lowest weight from $x_t^{w,s}$ is removed to obtain selection $x_t^{w,s-1}$. This process results in a set of selections $\mathcal{W} = \{x_t^{w,s}\}_{s=1}^{n}$.

In contrast to most of the earlier literature, I use an out-of-sample forecast design in this paper. I obtain variable selections from a pre-sample and determine the optimal selection size, i.e. find those selections in $\mathcal{L}$ and $\mathcal{W}$ that minimize the root mean squared error (RMSE) of predictions. This comes closer to application in real time and may reveal issues related to overfitting and spurious selections. Given the heuristic nature of variable selection, the two methods would in general result in different selections and the optimal selection sizes may differ. I therefore determine the optimal selection size separately for each method.

One difference between factor model prediction weights and LARS is that the former would select predictors with high commonality, while LARS would avoid strongly correlated predictors. To see this, consider a group of highly correlated predictors within x_t. From principcal components analysis, all elements of the group would attain similar factor loadings and therefore similar model

prediction weights. With LARS, by contrast, if one element of the group gets included in set $x_t^{L,s}$, the new residual will have a low correlation with the remaining elements of this group (Bai and Ng, 2008). Hence, the latter would no longer be selected. Overall, LARS is therefore likely to result in a more diverse final set of predictors than prediction weights. Arguably, as regards the estimation of factors, the intuition of stepwise regression is not consistent with the selection of optimal variables. It may be therefore more desirable to select variables with high commonality.

3 Prediction weights from a dynamic factor model

This section discusses prediction weights in the context of the dynamic factor model by Doz et al. (2011). The model is given by equation (1) together with the law of motion

$$f_{t+1} = \sum_{l=1}^{p} \Psi_l f_{t-l+1} + B\eta_t, \qquad \eta_t \sim \mathbb{N}(0, I_q). \tag{5}$$

Common factors f_t are driven by q-dimensional white noise η_t with $r \times q$ matrix B, where $q \leq r$. The stochastic process for f_t is assumed to be stationary. Further, the idiosyncratic component ξ_t is modelled from multivariate white noise with diagonal covariance matrix Σ_ξ. [2]

In the empirical application, I will use the factor model to predict quarterly GDP growth from monthly data x_t. To handle these mixed frequencies, I follow Harvey (1989: 309ff) and introduce monthly GDP growth y_t as a latent variable (see also e.g. Mariano and Murasawa, 2010; Angelini et al., 2011). y_t is assumed to be related to factors f_t by equation

$$y_t = \mu + \beta' f_t + \varepsilon_t \qquad \varepsilon_t \sim \mathbb{N}(0, \sigma_\varepsilon^2). \tag{6}$$

This is supplemented with log-linear aggregation rules to relate y_t to observed quarterly GDP growth, y_t^Q. For this purpose, another latent variable Q_t is defined at monthly frequency such that it corresponds to y_t^Q in the 3^{rd} month of the respective quarter, $t = 3k$. Aggregation rules

[2] Data x_t load only on current values of factors. However, the representation of Doz et al. (2011) can be derived from a version of a general DFM with q dynamic factors where x_t loads on current and lagged values (see Stock and Watson, 1995).

can then be expressed as

$$\begin{aligned} y_t^{(3)} &= y_t + y_{t-1} + y_{t-2} \\ Q_t &= \frac{1}{3}(y_t^{(3)} + y_{t-1}^{(3)} + y_{t-2}^{(3)}) \\ y_{3k}^{Q} &= Q_{3k}, \qquad k = 1, 2, ..., \lfloor T/3 \rfloor . \end{aligned} \tag{7}$$

where $y_t^{(3)}$ represents 3-month growth rates of monthly GDP, i.e. growth rates vis-a-vis the same month of the previous quarter. In application, y_t^Q is treated as missing in months 1 and 2 of the quarter, but added to observation vector z_t in month 3.

Equations (1), (5), (6), and the aggregation rules can be cast in a single state space form with state vector $\alpha_t = (f_t, \ldots, f_{t-p+1}, y_t, y_{t-1}, y_t^{(3)}, Q_t)$. The state space form is given in annex 1.

$$\begin{aligned} z_t &= W_t \alpha_t + u_t \qquad\qquad u_t \sim \mathbb{N}(0, \Sigma_u) \\ \alpha_{t+1} &= T_t \alpha_t + v_t, \qquad\qquad v_t \sim \mathbb{N}(0, \Sigma_v) \end{aligned} \tag{8}$$

The Kalman filter and associated smoothing algorithms (see e.g. Durbin and Koopman, 2001) provide minimum mean square linear (MMSE) estimates $a_{t+h|t} = \mathbb{E}[\alpha_{t+h}|\mathcal{Z}_t]$ of the state vector and their covariance $P_{t+h|h}$ for information set $\mathcal{Z}_t$ and any $h > -t$.

Estimation of the model parameters is described in Giannone et al. (2008). Briefly, estimates of factor loadings Λ and initial estimates of factors f_t are obtained from principal components. The latter are used to estimate Ψ_l in equation (5) from OLS. A further application of principal components to the residual covariance matrix of the VAR then gives matrix B. Parameters β and σ_ε^2 are estimated from a quarterly version of equation (6), again using the initial estimates of factors f_t with appropriate adjustments (see Angelini et al., 2011).

I will use information criteria to obtain the model specifications. Specifically, r, p, and q are found at the various stages of the estimation process from criterion PCP_2 in Bai and Ng (2002), the Akaike information criterion (AIC), and criterion 2 in Bai and Ng (2007), respectively.[3]

[3] Jungbacker et al. (2011) and Banbura and Modugno (2014) present maximum likelihood methods to estimate the model, possibly with missing data. They report gains in forecast precision to be limited. Given the high number of estimates in my experiments, with recursive estimation in a variable selection loop, I stick to the less time-consuming two-step estimator.

As pointed out by Bańbura and Rünstler (2011), prediction weights for y_t^Q can be obtained from an extension of the Kalman filter and smoother due to Harvey and Koopman (2003). For any infomation set $\mathcal{Z}_t$, the extension provides the weights of individual observations in estimates $a_{t+h|t}$ of the state vector, $h > -t$. As y_t^Q is an element of the state vector, this allows predictions $y_{t+h|t}^Q$ to be expressed as

$$y_{t+h|t}^Q = \sum_{l=0}^{t-1} \omega_{l,t}'(h) z_{t-l} \ . \tag{9}$$

with weights $\omega_{l,t}(h)$. Clearly, weights depend on both the forecast horizon h and the information set $\mathcal{Z}_t$. In recursive forecast evaluation exercises it is therefore important to define information sets such that the Kalman filter and smoothing algorithms approach their steady state and the time index on weights can be dropped. This holds for pseudo real-time data sets $\mathcal{Z}_t$ as defined below in section 5.

Since the Kalman filtering and smoothing algorithms provide MMSE estimates, weights $\omega_{i,l}(h)$ are a measure of the marginal predictive gain in $y_{t+h|t}^Q$ that arises from adding observation $x_{i,t-l}$ to the information set. In the below exercise, I will consider cumulative weights $\omega(h) = \sum_{l=0}^{k} \omega_l(h)$ as a measure of the predictive content of series $x_{i,t}$ for $y_{t+h|t}^Q$, where k is chosen sufficiently large.[4]

To obtain selections from LARS, the monthly data must be aggregated to quarterly frequency, x_t^Q. LARS selections for predictions $y_{t+h|t}^Q$ can then be obtained from static regressions of quarterly GDP growth y_t^Q on x_{t-h}^Q as from equation (3), $h = 0, 1, \ldots$.

4 A Monte Carlo simulation

This section conducts a Monte Carlo study to investigate the gains from the two variable selection methods. The simulation design is a variant of simulation 1 in Boivin and Ng (2006). I use the dynamic factor model described in section 3, but I abstract from mixed frequency issues and assume that y_t is observable.

[4] In application, weights $\omega_l(h)$ decay quickly unless factors f_t are highly persistent. The choice of k is therefore not critical. Cumulative weights do not measure the predictive gain of a series across all lags precisely. Such measure could be obtained from $P_{t+h|h}$ to find the loss in forecast precision when eliminating series j from the data (Giannone et al., 2008). However, this becomes computationally very expensive in a stepwise approach as it requires $O(n^2)$ runs of the Kalman filter and smoother.

The data are generated from the equations

$$\begin{aligned} x_t &= \lambda f_t + \xi_t, & \xi_t &\sim \mathbb{N}(0, \Sigma_\xi), \\ f_t &= \psi f_{t-1} + \eta_t, & \eta_t &\sim \mathbb{N}(0, \sigma_\eta^2), \\ y_t &= \beta' f_t + \varepsilon_t, & \varepsilon_t &\sim \mathbb{N}(0, \sigma_\varepsilon^2). \end{aligned}$$

I assume a single latent factor f_t, which is modelled as a first-order autoregressive process with $\sigma_\eta^2 = 1 - \psi^2$ such that $\text{var}(f_{i,t}) = 1$. The $n \times 1$ vector of series $x_t = (x_{1t}, \ldots, x_{nt})'$ and the scalar target series y_t are defined as in equations (1) and (2), while idiosyncratic component ξ_t is assumed to be multivariate white noise with covariance matrix Σ_ξ.

Factor loadings $\lambda = (\lambda_1, \ldots, \lambda_n)'$ are assumed to differ across series. They are drawn from a beta distribution $\mathbb{B}(a, b)$ over support $(0, 1)$. I will consider various values of a and b to inspect the role of dispersion and skewness of factor loadings on the success of variable selection methods. Heterogenous factor loadings translate into heteroscedasticity in idiosyncratic components, as series x_{it} are standardised to $\text{var}(x_{i,t}) = 1$, which implies $\text{var}(\xi_{i,t}) = 1 - \lambda_i^2$.

Further, I allow for non-zero cross correlations among idiosyncratic components ξ_{it}. I simply set $\text{corr}(\xi_{it}, \xi_{jt}) = \rho$ for all $i, j = 1, \ldots, n, \;\; i \neq j$. The elements ij of covariance matrix Σ_ξ are therefore given by

$$\Sigma_{\xi,ij} = \begin{cases} 1 - \lambda_i^2 & \text{for } i = j \\ \rho\sqrt{\left(1 - \lambda_i^2\right)\left(1 - \lambda_j^2\right)} & \text{otherwise} \end{cases}$$

The parameters of forecasting equation (2) for y_t are kept fixed with $\beta = 0.75$ and $\sigma_\varepsilon^2 = 1 - \beta^2$, which implies $\text{var}(y_t) = 1$.

As discussed in section 2, once $\rho > 0$, forecast performance may be improved from using a limited set of variables. With the above simulation design, because the correlations among idiosyncratic components are assumed to be identical across all series, the information content of series $x_{i,t}$ for estimating f_t depends only on its factor loading λ_i. Hence, the best selections would simply consist of the series with highest factor loadings.

The simulations aim at assessing the usefulness of LARS and factor model prediction weights for obtaining predictions $y_{t+h|t,}, \;\; h \geq 0$. I use an out-of-sample forecast design. I take 500 draws of

of length $T = 180$, which amounts to 15 years of monthly data. The number of series is set to $n = 100$. For each draw $\left\{f_{t,J}, \xi_{t,J}, \varepsilon_{t,J}, \lambda_J\right\}$ I obtain $\{x_{t,J}, y_{t,J}\}$ and proceed as follows:

1. I split draw J into two subsamples 1 and 2 of length $T_1 = T_2 = 90$.
2. From the pre-sample (sub-sample 1), I obtain variable selections of sizes $s = 1, \ldots, n$. Denote with $\mathcal{W}_J = \left\{x_{t,J}^{w,s}\right\}_{s=1}^{n}$ and $\mathcal{L}_J = \left\{x_{t,J}^{L,s}\right\}_{s=1}^{k}$ the sets of selections according to factor model prediction weights and LARS, respectively, as described in section 3. For obtaining prediction weights, I either keep the number of factors fixed at the true value of $r = 1$ or estimate r from information criterion PCP_2 in Bai and Ng (2002).

   I further determine the optimal number of series for either selection method from a minimum RMSE criterion. I choose the selection $x_{t,J}^{w,s}$ in $\mathcal{W}_J$, which gives the minimum root mean squared error (RMSE) of predictions $\widehat{y}_{t,J}^{w,s}$ in sub-sample 1. I proceed equivalently for $\mathcal{L}_J$.[5]
3. For all selections $x_{t,J}^{w,s}$ and $x_{t,J}^{L,s}$, I estimate the parameters of the dynamic factor model from sub-sample 1. I then obtain predictions $\widehat{y}_{t+h|t,J}^{w,s}$ and $\widehat{y}_{t+h|t,J}^{L,s}$ for y_t^J over both subsamples.

Table 1 shows the findings for the case of a static factor model, $\psi = 0$. The table reports the average RMSE of pre-sample and out-of-sample predictions, the number of series chosen by the in-sample RMSE criterion, and the precentage of correct classification. Denote with $x_{t,J}^{*,s}$ the $s \times 1$ vector of series with the highest s factor loadings among $x_{t,J}$. The percentage of correct classifications is given by the share of elements of $x_{t,J}^{*,s}$ contained in selections $x_{t,J}^{w,s}$ and $x_{t,J}^{L,s}$, respectively.

The following conclusions emerge from Table 1. First, for $\rho = 0$, prediction weights and LARS choose 54 and 20 series, respectively, as opposed to the optimal choice of 100. The smaller selections result in some small losses in out-of-sample predictions, as to be expected.

Second, once $\rho > 0$, both selection methods results in improved out-of-sample predictions. For some of the simulations, these gains are of considerable size. While losses against predictions based on the full set of series may occur, they are always minor. The optimal selections are

[5] For LARS I use code by Karl Skoglund (http://www.cad.zju.edu.cn/home/dengcai/Data/code/lars.m).

generally small: with one exception, the in-sample RMSE criterion chooses less than 20 series. Optimal selections from prediction weights are somewhat larger than those from LARS.

Third, factor model prediction weights consistently outperform LARS. They provide a lower out-of-sample RMSE, although often only by a small margin, and the percentage of correctly classified series is considerably higher. Whereas the share of correctly classified series amounts to about 0.7 to 0.8 in prediction weight selections, it is less than 0.5 in LARS selections. The overlap among the selections turns out to be moderate. In general, the share of series that are contained in both selections is in between 0.5 and 0.6, depending on the sizes of the selections.[6]

Perhaps more important, LARS shows some tendency of overfitting. For prediction weights, the gains indicated by in-sample predictions largely carry over to the out-of-sample case. The comparatively large pre-sample gains from LARS selections, however, turn out to be spurious, as gains in out-of-sample predictions are much smaller. This is most apparent for the case of $\rho = 0$, where applying either variable selection method results in a slight deterioration of the out-of-sample RMSE, as suggested by asymptotic theory. The pre-sample RMSE suggests however considerable gains from LARS selections.

Fourth, as to the role of ρ and (a, b), higher values of ρ and and skewed distributions of factor loadings with a high share of uninformative series give rise to larger gains from variable selection. Beta distribution $\mathbb{B}(a, b)$ is symmetric for $a = b$ with mean 0.5 and its dispersion declines with higher a and b. The case of $a = 1$, $b = 3$ amounts a to left-skewed distribution with mean 0.25, implying a high share of series with low factor loadings, while the opposite case of $a = 3$, $b = 1$ gives a right-skewed distribution with mean 0.75. The gains from variable selection raise with high dispersion and with a left-skewed distribution.

Fifth, for $\rho > 0$ the specific correlation structure of ξ_t in this exercise implies a single high eigenvalue of Σ_ξ and therefore the presence of one principal component in ξ_t. Once the number of factors r is estimated, this is occasionally picked up as a second factor. It turns out, that selection methods then act as an insurance against misspecification. The right hand lower panel

[6] Predictions based on based $x_{t,J}^{w,s}$ fall only marginally short of predictions based on $x_{t,J}^{*,s}$, i.e. under the assumption of perfect knowledge about the ranking of series.

Table 1: Monte Carlo Simulations

Static factor model

		Symmetric distributions (r = 1)						
Factor persistence ψ		0.0	0.0	0.0	0.0	0.0	0.0	0.0
Correlation (ρ)		0.0	0.1	0.1	0.1	0.3	0.3	0.3
a		4	2	4	8	2	4	8
b		4	2	4	8	2	4	8
Nr of factors (r)		1	1	1	1	1	1	1
Nr of series selected	Weights	53.5	16.1	19.7	26.9	7.6	9.9	13.1
	LARS	21.4	11.9	12.6	14.4	6.6	6.7	7.3
Correct classification	Weights	0.88	0.88	0.82	0.75	0.86	0.79	0.69
	LARS	0.35	0.36	0.38	0.37	0.45	0.45	0.41
RMSE pre-sample	All	0.66	0.70	0.72	0.74	0.78	0.80	0.82
	Weights	0.66	0.67	0.69	0.71	0.69	0.72	0.75
	LARS	0.59	0.63	0.64	0.64	0.67	0.70	0.72
RMSE out-of-sample	All	0.68	0.71	0.74	0.75	0.79	0.82	0.83
	Weights	0.68	0.69	0.71	0.74	0.70	0.74	0.78
	LARS	0.69	0.70	0.73	0.75	0.72	0.76	0.80

		Asymmetric distributions (r = 1)				Number of factors estimated			
Factor persistence ψ		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Correlation (ρ)		0.1	0.3	0.1	0.3	0.3	0.3	0.3	0.3
a		1	1	3	3	2	4	1	3
b		3	3	1	1	2	4	3	1
Nr of factors (r)	All	1	1	1	1	1.99	1.14	1.48	2.00
	Weights	1	1	1	1	1.55	1.01	1.07	1.74
	LARS	1	1	1	1	1.72	1.01	1.09	1.79
Nr of series selected	Weights	12.7	6.8	26.3	13.9	33.5	19.1	13.9	29.7
	LARS	10.3	6.2	11.9	8.0	22.7	12.5	10.2	16.3
Correct classification	Weights	0.81	0.77	0.94	0.92	0.70	0.67	0.69	0.75
	LARS	0.49	0.52	0.18	0.24	0.43	0.45	0.51	0.24
RMSE pre-sample	All	0.82	0.92	0.66	0.69	0.67	0.80	0.81	0.66
	Weights	0.72	0.77	0.65	0.66	0.67	0.73	0.78	0.65
	LARS	0.69	0.77	0.62	0.64	0.64	0.70	0.78	0.61
RMSE out-of-sample	All	0.84	0.94	0.67	0.70	0.69	0.82	0.84	0.68
	Weights	0.75	0.79	0.67	0.67	0.70	0.76	0.81	0.68
	LARS	0.76	0.81	0.68	0.69	0.73	0.78	0.84	0.70

The table shows findings from Monte Carlo simulations for the static factor model (ψ=0). ρ is the correlation among idiosyncratic components, while a and b are the parameters of the beta distribution, from which factor loadings λ are drawn (see main text). The table shows four statistics for both LARS and prediction weights: (i) the average number of series in the optimal selections (selections with minimum RMSE pre-sample); (ii) the percentage of correct classifications, i.e. of series with highest factor loadings; (iii) the RMSE of the optimal selection in the pre-sample; column 'All' refers to the RMSE from all 100 series; and (iv) the out-of-sample RMSE of the optimal selection. The upper panel shows results for symmetric distributions of factor loadings with the number of factors kept fixed at r = 1. The lower left-hand panel shows results for asymmetric distributions with r = 1. The lower right hand panel shows results where the number of factors is estimated from an information criterion (see main text).

Table 2: Monte Carlo Simulations

Dynamic factor model

		Static predictions (h = 0)							
Factor persistence ψ		0.5	0.5	0.5	0.5	0.8	0.8	0.8	0.8
Correlation (ρ)		0.1	0.1	0.3	0.3	0.1	0.1	0.3	0.3
a		4	8	4	8	4	8	4	8
b		4	8	4	8	4	8	4	8
Nr of factors (r)		1	1	1	1	1	1	1	1
Nr of series selected	Weights	20.2	25.6	9.4	12.4	19.2	24.5	9.5	11.8
	LARS	12.7	14.6	6.6	7.4	12.4	13.9	6.5	7.2
Correct classification	Weights	0.82	0.74	0.79	0.68	0.81	0.71	0.78	0.65
	LARS	0.38	0.36	0.45	0.40	0.37	0.35	0.44	0.40
RMSE pre-sample	All	0.72	0.74	0.81	0.82	0.72	0.74	0.80	0.82
	Weights	0.69	0.71	0.72	0.76	0.69	0.71	0.72	0.76
	LARS	0.64	0.64	0.70	0.72	0.64	0.64	0.70	0.72
RMSE out-of-sample	All	0.74	0.75	0.82	0.83	0.74	0.75	0.82	0.83
	Weights	0.71	0.74	0.74	0.78	0.71	0.74	0.74	0.78
	LARS	0.73	0.75	0.76	0.79	0.72	0.75	0.76	0.79

		One-step ahead predictions (h=1)							
Factor persistence ψ		0.5	0.5	0.5	0.5	0.8	0.8	0.8	0.8
Correlation (ρ)		0.1	0.1	0.3	0.3	0.1	0.1	0.3	0.3
a		4	8	4	8	4	8	4	8
b		4	8	4	8	4	8	4	8
Nr of factors (r)		1	1	1	1	1	1	1	1
Nr of series selected	Weights	26.8	30.2	13.8	15.2	19.3	27.0	9.6	12.8
	LARS	25.1	27.5	11.2	10.7	17.4	20.4	6.7	7.7
Correct classification	Weights	0.83	0.75	0.79	0.69	0.82	0.74	0.78	0.67
	LARS	0.47	0.44	0.47	0.42	0.43	0.41	0.46	0.40
RMSE pre-sample	All	0.93	0.94	0.96	0.96	0.82	0.83	0.89	0.89
	Weights	0.92	0.93	0.93	0.94	0.79	0.80	0.81	0.85
	LARS	0.92	0.93	0.94	0.94	0.79	0.80	0.82	0.84
RMSE out-of-sample	All	0.95	0.95	0.97	0.98	0.84	0.85	0.91	0.91
	Weights	0.94	0.95	0.95	0.96	0.81	0.83	0.84	0.87
	LARS	0.95	0.95	0.96	0.96	0.82	0.84	0.85	0.88

See Table 1 for a description of the contents.

of Table 1 shows simulation results, where the number of factors r is estimated from information criterion PCP_2 (Bai and Ng, 2002). For $\rho = 0.1$, the criterion chooses $r = 1$ in all cases and gains from variable selection prevail. For $\rho = 0.3$, this still holds for some of the values of (a, b) considered in the simulations. There arise yet two cases where r is estimated predominantly with 2. Predictions from $r = 2$ and the full data set then perform as well as the optimal predictions from $r = 1$, while selection methods do not deliver further gains. In this case, hence, variable selection acts to avoid the losses that would arise from choosing $r = 1$, i.e. lower as suggested by the in-sample information criterion. Note that this pattern occurs for precisely those simulations, where selection methods had delivered the largest gains under $r = 1$.

Finally, Table 2 shows that the above conclusions for the case of $\psi = 0$ straightforwardly carry over to $\psi > 0$. Results are reported for both static predictions $y_{t|t}$ and one-step ahead forecasts $y_{t+1|t}$. For static predictions, the above findings remain almost unchanged for both $\psi = 0.5$ and $\psi = 0.8$. For one-step ahead predictions, the gains from variable selection remain in case of highly persistent factor dynamics ($\psi = 0.8$). However, for $\psi = 0.5$, gains decline considerably, as the predictions become generally less informative.

5 Forecasting GDP growth from monthly data

This section presents a pseudo real-time exercise to obtain now- and forecasts of quarterly GDP growth in the euro area, Germany, and France from large unbalanced monthly data sets. I obtain variable selections based on factor model prediction weights and LARS from a pre-sample and evaluate their performance from a pseudo real-time forecast exercise in the second part of the sample.

The dynamic factor model by Doz et al. (2011) has been applied to predict quarterly GDP growth from unbalanced monthly data for a number of countries, including the U.S. (Giannone et al., 2008), the euro area (Angelini et al., 2011), and several euro area member states, such as Germany and France (Rünstler et al., 2009; Marcellino and Schumacher, 2010; Schumacher, 2010; Bessec, 2013). Rünstler et al. (2009) and Marcellino and Schumacher (2010) find the model to

perform about as well as other versions of dynamic factor models, while Rünstler et al. (2009) and Angelini et al. (2011) report that it is superior to pooled forecasts from single equations.

I use monthly data sets for the euro area, Germany, and France of each about 70 series. All data start in January 1991. They were downloaded on 26, Sep 2014. The choice of series is based on Angelini et al. (2011) and includes data on economic activity (such as industrial production, trade, employment), the European Commission business and consumer surveys, financial markets, and the international environment. The series are transformed to monthly rates of change and standardised to mean zero and variance one. Further, they are cleaned from outliers. The series are listed in annex A together with their publication lags and the data transformations used.[7]

5.1 Forecast design

The forecast design follows Angelini et al. (2011) and aims at replicating the real-time application of the factor model as closely as possible.

First, I account for the timing of data releases. Real-time data sets typically contain missing observations at the end of the sample due to publication lags. Survey and financial market data, for instance, are available right at the end of the respective month, while data on economic activity are usually published with a delay of 6 to 8 weeks. Giannone et al. (2008) and Bańbura and Rünstler (2011) report that differences in the timing of data releases among individual series have large effects on their marginal predictive gains.

I therefore follow those studies in applying so-called *pseudo-real time* data sets $\mathcal{Z}_t$, which employ the final data release, but replicate the publication lags from the end of the sample in the earlier periods. Let $z_t' = (x_t', y_t^Q)$ and denote with $\mathcal{Z}_t$ the information set in period t. Consider the original data set $\mathcal{Z}_T$ as downloaded in period T. Data set $\mathcal{Z}_t$, on which the forecast in period t is based, is obtained by eliminating observation $x_{i,t-l}$, $l \geq 0$, if and only if observation $x_{i,T-l}$ is missing in $\mathcal{Z}_T$, $i = 1, \ldots, n$. Quarterly GDP growth is treated in an equivalent way. Kalman

[7] Outliers are defined as observations that deviate by more than twice the interquintile distance from the median. The interquintile distance is defined as the difference between the 80% and 20% quantiles of the empirical distribution. For principal components, outliers are replaced with the median, for Kalman filtering they are set as missing.

filtering and smoothing handles unbalanced data sets in an efficient way. The rows in equation (8) corresponding to missing observations in z_t are simply skipped when applying the respective recursions (Durbin and Koopman, 2001:92f).

Second, I inspect six predictions for GDP growth in a certain quarter, which are obtained in consecutive months. I start in the 1^{st} month of the previous quarter and stop in the 3^{rd} month of the current quarter, 6 weeks before the flash estimate of GDP is released. To predict GDP growth in the 2^{nd} quarter, for instance, the 1^{st} prediction is run in January and the final (6^{th}) one in July. Note that predictions 4 to 6 amount to nowcasting the current quarter.

Third, I inspect out-of sample predictions. I obtain the variable selections and corresponding factor model specifications from a pre-sample and run a forecast exercise with recursive parameter estimation on the remainder. I proceed as follows:

1. I obtain selections $\{x_t^{w,s}\}_{s=1}^{n}$ and $\left\{x_t^{L,s}\right\}_{s=1}^{k}$ from the pre-sample ranging until 2000 Q4 using stepwise elimination as described in section 2.1. I use different selections for now- and next-quarter forecasts. Prediction weight selections are based on mid-quarter weights, i.e. weights from prediction 5 for nowcasts (predictions $4-6$) and 2 for the next-quarter forecasts (predictions $1-3$), respectively.

   While factor model prediction weights account for publication lags, a standard application of LARS would ignore them. Bessec (2013) argues that LARS selections can be improved by accounting for publication lags. She proposes to start with unbalanced data and to forecast the missing observations from univariate methods. I follow a proposal by Altissimo et al. (2010) instead and shift the monthly series prior to aggregation.

   That is, with series $x_{i,t}$ being subject to a publication lag of l months, I define $x_{i,t}^{\#} = x_{i,t-l}$. I run LARS with quarterly GDP growth being regressed on quarterly aggregates $x_{i,t}^{\#Q}$ at either lag 0 for nowcasts or 1 next-quarter predictions.

   For all selections, model specifications are obtained from the information criteria set out in section 3. The model is re-specified at each selection step under the restriction that the dimensionality of the model shrinks with the number of series. That is, e.g., for prediction

weights I obtain specifications $(r^{w,s}, p^{w,s}, q^{w,s})$ related to selection $x_t^{w,s}$ under the restrictions $r^{w,s} \leq r^{w,s+1}$, $p^{w,s} \leq p^{w,s+1}$, and $q^{w,s} \leq q^{w,s+1}$ for $s < n$.

2. I obtain now- and next-quarter forecasts of GDP growth for the period starting with 2001 Q1 based on the variable selections as from step 1. These forecasts employ pseudo real-time data sets $\mathcal{Z}_t$ and recursive parameter estimates.

   The financial crisis requires some special consideration in the choice of the evaluation sample. It not only implies extremely large forecast errors in 2008 and 2009, but may also constitute a structural break in economic activity in the euro area. The evaluation of variable selection methods may therefore be more safely based on the pre-crisis period. On the other hand, the performance of the models after the crisis is certainly of interest.

   I therefore evaluate the forecasts separately for two samples, a pre-crisis sample from 2001 Q1 to 2007 Q4 and a post-crisis sample ranging from 2010 Q1 to 2014 Q2.

5.2 Results

The factor model specifications chosen by the information criteria are similar across data sets. For the euro area and German full data, the number of factors is estimated with $r = 3$, while p and q are estimated with 2. For the various selections, estimates of r remain at 3, while estimates of p and q shrink as the number of series declines. For France, r and q shrink from 3 to 1, while p stays at 3. For all countries, the average cross correlation among idiosyncratic components is slightly below 0.2. Idiosyncratic components are subject to considerable heteroscedasticity.

Table 3 shows the RMSE of predictions for the sequence of 6 predictions described above. The numbers are averaged over predictions $1-3$ (next-quarter) and $4-6$ (nowcasts). The RMSE is shown relative to the naive forecast, which is the sample mean of GDP growth.[8] Note that variable selection with LARS stops at $s = 30$, as it is limited by the number of observations.

Starting with predictions from the full data set, for the pre-sample and the pre-crisis evaluation sample, the factor model predictions in general improve upon the naive forecast and a first-order

[8] The calculation of naive and AR(1) forecasts takes account of the timing of the publication dates of the GDP flash estimates. Forecasts are based on recursive estimates.

autoregression for GDP ($AR(1)$). The small gains against the $AR(1)$ in the euro area pre-crisis nowcasts fall somewhat short of the findings reported by Angelini et al. (2011), obtained from a shorter sample. The results for Germany and France are largely in line with earlier studies (e.g. Rünstler et al., 2009; Schumacher, 2010; Barhoumi et al., 2010). For the post-crisis sample, the performance of the factor model worsens for the euro area and France, with predictions being outperformed by the $AR(1)$.

The ranking of the series according to the selections from nowcasts are shown in Tables A.1 to A.3 in the annex. Selections from prediction weights are less heterogenous than those from LARS and there is very little overlap among the two. For the euro area, prediction weight selections contain the main items of business surveys (confidence indicators and order books), together with equity price indices, the euro area real effective exchange rate, and raw materials prices. For Germany and France, business, consumer and construction survey items are prominent. Conversely, LARS puts more weight on hard data, such as items of industrial production, and items of construction and retail trade surveys. As discussed in section 2, LARS also tends to select a more diverse set of series.

I turn to the performance of variable selections. For the euro area, both methods improve predictions $4-6$ (nowcasts), but have little effect on predictions $1-3$ (next-quarter forecasts). Prediction weight and LARS selections of $10-15$ and $20-30$ series, respectively, perform best, with gains of about 15% compared to the full data set. Conversely, for predictions $1-3$ there are no gains from either method, with small selections of 20 series or less giving rise to sizeable losses. Crucially, these patterns are properly detected in the pre-sample. In real-time application, both methods would therefore have chosen the correct selections.

Results are more mixed for Germany and France. For Germany, the pre-sample indicates moderate gains of less than 10% in nowcasts from small selections of $10-15$ series from either method. For factor model prediction weights, these gains carry over to the pre-crisis sample and vanish in the post-crisis one. The application of LARS selections would however result in losses in either sample. The same applies to LARS selections for next-quarter predictions, although the gains indicated in the pre-sample are very small.

For France, the pre-sample indicates gains from both selection methods over the entire horizon. Small gains from factor model prediction weights arise for selections of $40 - 60$. Again, these gains carry over to both evaluation samples. The gains from LARS selections are sizeable in the pre-sample. In evaluation samples however, they realize only in the post-crisis samples, whereas in the pre-crisis sample LARS selections give rise to losses at all horizons.

These findings are summarized in Fig. 1, which compares the RMSEs from the full data set with those of the factor model prediction weight and LARS selections that are found to give the smallest RMSE in the pre-sample. Generally, prediction weights appear more robust than LARS selections. They give rise to modest, but stable gains in nowcasts over a range of selection sizes and the pre-sample gives largely correct signals on appropriate selections. While some selections might give rise to losses in next-quarter predictions, these are properly detected in the pre-sample. LARS selections fare equally well for the euro area, but give more mixed results for Germany and France. In particular, the pre-sample gives wrong signals for out-of sample predictions in Germany over the entire horizon, and the pre-crisis sample in France.[9]

The above conclusions withstand various robustness checks. First, I used fixed model specifications, i.e. applied the specification (r, p, q) as obtained from the full data set to all variable selections. Second, I inspected whether the selections obtained from nowcasts (i.e. prediction 5) may help in improving next-quarter predictions. Third, I used LARS selections derived from the original data $x_{i,t}$ instead of shifted data $x_{i,t}^{\#}$ as described in section 5.1. These modifications had overall small effects on the results. As one exception, LARS selections for German next-quarter predictions were found to be uninformative in the pre-sample, which avoids the losses in the corresponding predictions in evaluation samples.

[9] I do not present tests for forecast accuracy, as they are computationally very costly. The tests give rise to non-standard test distributions, because the individual selections are nested in the full data set, which requires bootstrap techniques. While Hubrich and West (2010) provide a test statistic for nested models that uses standard distributions, the test is not applicable as the (one-sided) alternative hypothesis goes in the wrong direction: the test examines whether *adding* data to a minimal model would help in reducing forecast errors.

Table 3: Forecasting performance of selections

Euro area

	Naive	AR(1)	All	Prediction weights							LARS			
				60	50	40	30	20	15	10	30	20	15	10
Pre-sample (1991 Q1 - 2000 Q4)														
4-6	0.488	0.93	0.91	0.79	0.84	0.87	0.82	0.81	0.76	0.77	0.81	0.86	0.88	0.89
1-3	0.488	0.99	0.80	0.80	0.80	0.80	0.81	0.84	0.87	0.90	0.88	0.92	0.93	0.99
1-7	0.488	0.96	0.86	0.80	0.82	0.83	0.81	0.83	0.82	0.83	0.85	0.89	0.90	0.94
Pre-crisis (2001 Q1 - 2007 Q4)														
4-6	0.339	0.86	0.86	0.85	0.75	0.75	0.75	0.77	0.77	0.78	0.77	0.85	0.82	0.85
1-3	0.344	0.96	0.79	0.79	0.80	0.80	0.81	0.85	0.82	0.84	0.86	0.94	0.94	0.94
1-7	0.342	0.91	0.82	0.82	0.77	0.78	0.78	0.81	0.80	0.81	0.82	0.90	0.88	0.90
Post-crisis (2010 Q1 - 2014 Q2)														
4-6	0.441	0.76	1.01	0.99	0.88	0.87	0.83	0.79	0.80	0.76	0.82	0.64	0.65	0.72
1-3	0.444	0.85	0.89	0.92	0.91	0.92	0.95	0.95	0.96	0.97	0.83	0.92	0.92	0.93
1-7	0.442	0.80	0.95	0.96	0.89	0.89	0.89	0.87	0.88	0.86	0.83	0.78	0.79	0.82

Germany

	Naive	AR(1)	All	Prediction weights							LARS			
				60	50	40	30	20	15	10	30	20	15	10
Pre-sample (1991 Q1 - 2000 Q4)														
4-6	0.712	1.00	0.95	0.95	0.93	0.96	1.10	0.95	0.92	0.92	0.95	0.95	0.94	0.86
1-3	0.712	1.00	0.93	0.91	0.92	0.91	0.92	0.92	0.91	0.92	0.90	0.91	0.91	0.92
1-6	0.712	1.00	0.94	0.93	0.92	0.94	1.01	0.94	0.92	0.92	0.93	0.94	0.94	0.89
Pre-crisis (2001 Q1 - 2007 Q4)														
4-6	0.565	1.00	0.89	0.89	0.89	0.90	0.89	0.89	0.84	0.84	0.95	0.89	0.90	0.92
1-3	0.569	1.00	0.91	0.91	0.91	0.91	0.92	0.92	0.91	0.91	0.97	0.97	0.97	0.96
1-6	0.567	1.00	0.90	0.90	0.90	0.91	0.90	0.91	0.88	0.88	0.96	0.93	0.94	0.94
Post-crisis (2010 Q1 - 2014 Q2)														
4-6	0.633	0.96	0.82	0.83	0.77	0.89	0.83	0.83	0.81	0.81	0.86	0.86	0.87	0.89
1-3	0.635	0.97	0.88	0.86	0.86	0.85	0.85	0.85	0.91	0.95	1.00	1.02	1.03	0.99
1-6	0.634	0.97	0.85	0.85	0.81	0.87	0.84	0.84	0.86	0.88	0.93	0.94	0.95	0.94

France

	Naive	AR(1)	All	Prediction weights							LARS			
				60	50	40	30	20	15	10	30	20	15	10
Pre-sample (1991 Q1 - 2000 Q4)														
4-6	0.446	0.80	0.82	0.80	0.80	0.80	0.83	0.83	0.84	0.83	0.73	0.76	0.76	0.79
1-3	0.446	0.91	0.81	0.82	0.82	0.81	0.82	0.88	0.97	1.05	0.77	0.76	0.76	0.76
1-6	0.446	0.85	0.82	0.81	0.81	0.80	0.83	0.86	0.90	0.94	0.75	0.80	0.80	0.83
Pre-crisis (2001 Q1 - 2007 Q4)														
4-6	0.339	1.00	0.80	0.75	0.72	0.76	0.76	0.77	0.78	0.80	0.87	0.93	0.94	0.98
1-3	0.342	0.98	0.87	0.83	0.84	0.85	0.86	0.91	0.90	0.97	0.90	0.91	0.93	0.94
1-6	0.340	0.99	0.84	0.79	0.78	0.80	0.81	0.84	0.84	0.89	0.89	0.92	0.94	0.96
Post-crisis (2010 Q1 - 2014 Q2)														
4-6	0.388	1.00	1.02	0.90	0.89	1.03	1.04	1.07	1.07	1.05	0.83	0.81	0.81	0.77
1-3	0.389	0.95	1.08	1.02	0.98	0.99	0.99	1.37	1.26	0.90	0.96	0.93	0.92	0.91
1-6	0.389	0.98	1.05	0.96	0.94	1.01	1.02	1.22	1.17	0.98	0.90	0.87	0.87	0.84

Column 1 shows the RMSE of the naive forecast, based on a random walk with drift. The remaining columns show the RMSE relative to the naive forecast for an autoregressive model (AR(1)), the factor model with the full set of series (All), and various selections of different sizes from prediction weights and from LARS. The individual rows show the relative average RMSE over predictions 1 to 3 (next-quarter forecasts) 4 to 6 (nowcasts), and 1 to 6 (overall average), respectively. Results are shown for three separate sub-samples, i.e. the pre-sample, used fo variable selection, and pre- and post-crisis evaluation samples.

Figure 1: RMSE from best pre-sample selections

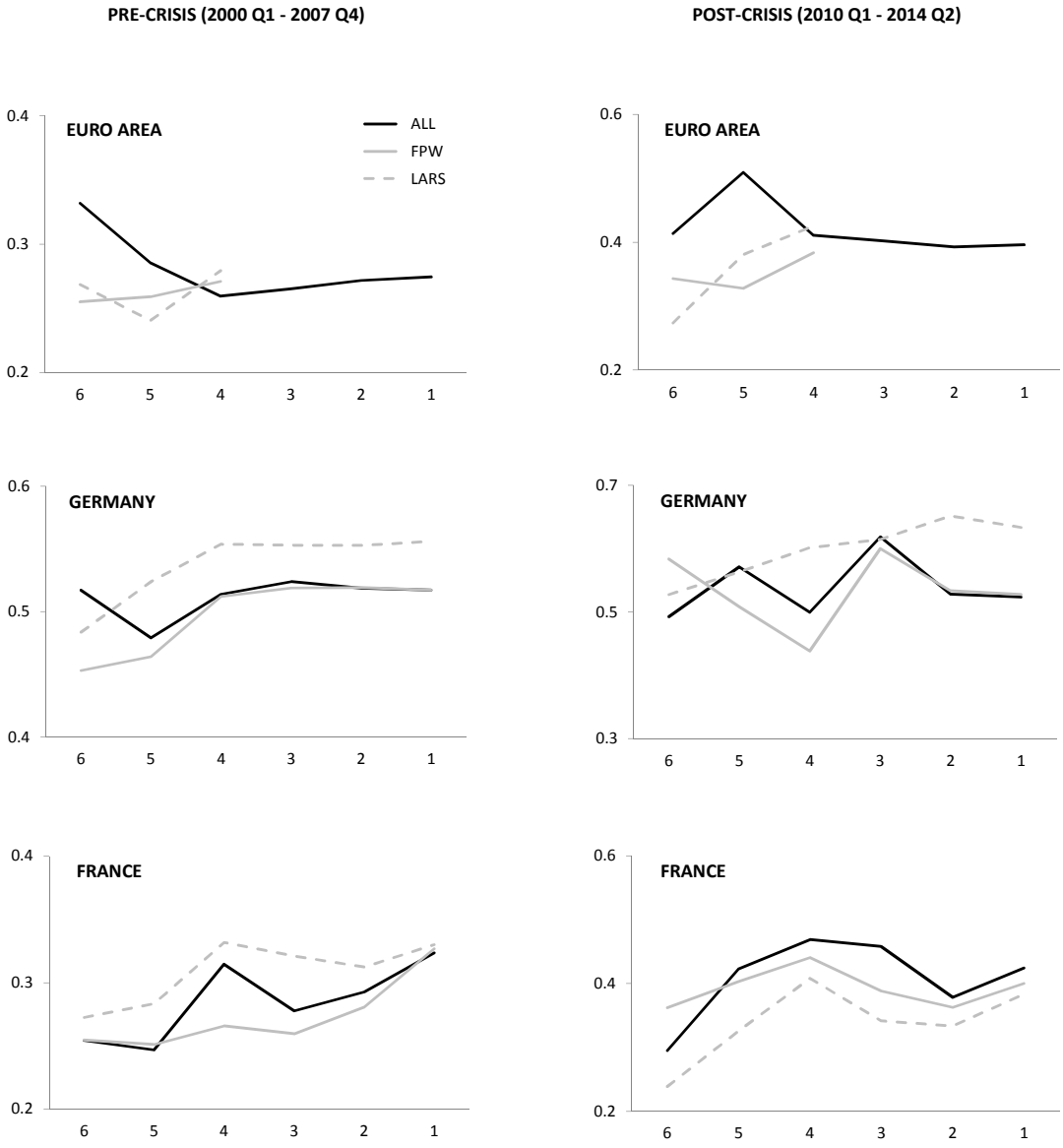


Figure 1 shows the out-of-sample root mean squared error (vertical axes) of predictions 1 to 6 (horizontal axis) from the optimal selections. ALL refers to predictions based on all series, whereas FPW and LARS refer to selections from factor model prediction weights and LARS, respectively. The optimal selections have been determined from the pre-sample. The left-hand and right-hand panels show the RMSE in the pre- and post-crisis evaluation samples, respectively. For the euro area predictions 1 to 3, the optimal selection is given by the full set of series (ALL).

6 Conclusions

The paper has inspected the efficiency gains from variable selection in predictions from a dynamic factor model. I have compared two methods for this purpose, i.e. Least Angle Regressions (LARS) and factor model prediction weights.

Against earlier studies by Bai and Ng (2008), Schumacher (2010) and Caggianio et al. (2011), which performed variable selection in the evaluation sample, this paper inspects the success of variable selection from a pre-sample. The results still confirm the earlier findings that variable selection methods tend to improve the efficiency of predictions. However, gains are moderate and should not be taken for granted. First, both the Monte Carlo simulations and the empirical findings indicate that such gains are small, at best, for one-step ahead forecasts. Second, the Monte Carlo simulations suggest that the relationship between the specification of the factor model and the success of variable selection is not straightforward.

For these reasons, variable selection methods should, first of all, be robust against avoiding potential losses in forecast precision in an out-of-sample context. The evidence presented in this paper suggests that factor model prediction weights perform better than LARS in this respect. In the Monte Carlo simulations they were better in identifying informative series and provided smaller out-of-sample forecast errors. LARS, in turn, showed signs of overfitting: pre-sample forecasts suggested gains that did not necessarily carry over to the out-of-sample case. Similarly, in the empirical application, pre-sample selections from LARS occasionally gave wrong signals that resulted in losses in out-of-sample predictions, whereas prediction weights provided consistent gains.

In the context of a dynamic factor model, factor model prediction weights obviously provide a model-consistent means of variable selection. One question for future research is whether they are useful for the pre-screening of variables also in the context of other forecasting methods.

References

Altissimo, F., R. Cristadoro, M. Forni, M. Lippi, and G. Veronese, 2010, New EuroCoin: tracking economic growth in real time, The Review of Economics and Statistics, 92(4): 1024–1034.

Angelini, E., G. Camba-Mendez, D. Giannone, L. Reichlin and G. Rünstler, 2011, Short-term forecasts of euro area GDP growth, Econometrics Journal, 14, C25-C44.

Alvarez, R., M. Camacho, and G. Perez-Quiros, 2012, Finite sample performance of small versus large scale dynamic factor models, CEPR discussion paper 8867.

Bai, J. and S. Ng, 2002, Determining the number of factors in approximate factor models, Econometrica, 70(1), 191-221.

Bai, J. and S. Ng, 2007, Determining the number of primitive shocks in factor models, Journal of Business and Economics Statistics, 25, 52-60.

Bai, J. and S. Ng, 2008, Forecasting economic series using targeted predictors, Journal of Econometrics 146, 304-317.

Bańbura, M. and M. Modugno, 2014, Maximum likelihood estimation of factor models on data sets with arbitrary pattern of missing data, Journal of Applied Econometrics 29(1), 133-160.

Bańbura, M. and G. Rünstler, 2011, A look into the factor model black box: publication lags and the role of hard and soft data in forecasting GDP, International Journal of Forecasting, 27(2), 333-346.

Barhoumi, K., O. Darné, and L. Ferrara, 2010, Are disaggregate data useful for factor analysis in forecasting French GDP?, Journal of Forecasting 29(1-2), 132-144.

Bessec, M., 2013, Short-term forecasts of French GDP: a dynamic factor model with targeted predictors, Journal of Forecasting 32, 500-511.

Boivin, J. and S. Ng, 2006, Are more data always better for factor analysis?, Journal of Econometrics 132(1), 169-194.

Caggiano, G., G. Kapetianos and V. Labhard, 2011, Are more data always better for factor analysis: results for the euro area, the six largest euro area countries and the UK, Journal of Forecasting 30, 736-752.

Doz, C., D. Giannone, and L. Reichlin, 2011, A quasi maximum likelihood approach for large approximate dynamic factor models, Journal of Econometrics, 164(1), 188-205.

Durbin, J. and S.J. Koopman, 2001, Time Series Analysis By State Space Methods, Oxford University Press.

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, 2004, Least angle regression, Annals of Statistics 32(2), 407-499.

Giannone, D., L. Reichlin, and D. Small, 2008, Nowcasting: the real-time informational content of macroeconomic data, Journal of Monetary Economics, 55(4), 665-676.

Harvey, A.C., 1989, Forecasting, Structural Time Series Models, and the Kalman filter, Cambridge University Press.

Harvey, A.C. and S.J. Koopman, 2003, Computing observation weights for signal extraction and filtering, Journal of Economic Dynamics and Control, 27, 1317-1333.

Hubrich, K. and K. West, 2010, Forecast evaluation of small nested model sets, Journal of Applied Econometrics 25(4), 574-594.

Jungbacker, B., S.J. Koopman, and M. van der Wel, 2011, Dynamic factor analysis in the presence of missing data, Journal of Economic Dynamics and Control 35(8), 1358 -1368.

Marcellino, M. and C. Schuhmacher, 2010, Factor-MIDAS for now- and forecasting with ragged-edge data: a model comparison for German GDP, Oxford Bulletin of Economics and Statistics 72, 518-550.

Mariano, R. and Y. Murasawa, 2010, A coincident index, common factors, and monthly real GDP, Oxford Bulletin of Economics and Statistics 72(1), 27-46.

Poncela, P. and E. Ruiz, 2015, More is not always better: back to the Kalman filter in dynamic factor models, in Koopman, S.J and N.G. Shephard (eds.), Unobserved Components and Time Series Econometrics, Oxford University Press.

Rünstler, G., K. Barhoumi, S. Benk, R. Cristadoro, A. den Reijer, A. Jakataine, P. Jelonek, A. Rua, K. Ruth, C. van Nieuwenhuyze, 2009, Short-term forecasting of GDP using large data sets, Journal of Forecasting 28(7), 595-611.

Schumacher, C., 2010, Factor forecasting using international targeted predictors: the case of German GDP, Economics letters 107(2), 95-98.

Stock, J.H. and M.W. Watson, 1995, Implications of dynamic factor models for VAR analysis, Princeton University mimeo.

Stock, J. H. and M. W. Watson, 2002a, Macroeconomic forecasting using diffusion indexes, Journal of Business and Economics Statistics, 20, 147-162.

Stock, J.H. and M.W. Watson, 2002b, Forecasting using principal components from a larger number of predictors, Journal of the American Statistical Association 97, 1167-1179.

State space form

The transition equation of the model described in section 3.1 with $p = 1$ is given by

$$\begin{bmatrix} I_r & 0 & 0 & 0 & 0 \\ -\beta' & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & -1 & 1 & 0 \\ 0 & 0 & 0 & -\frac{1}{3} & 1 \end{bmatrix} \begin{bmatrix} f_{t+1} \\ y_{t+1} \\ y_t \\ y_{t+1}^{(3)} \\ Q_{t+1} \end{bmatrix} = \begin{bmatrix} 0 \\ \mu \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \Psi_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \Xi_t \end{bmatrix} \begin{bmatrix} f_t \\ y_t \\ y_{t-1} \\ y_t^{(3)} \\ Q_t \end{bmatrix} + \begin{bmatrix} B\eta_t \\ \varepsilon_{t+1} \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

where I_r denotes the $r \times r$ identity matrix. Temporal aggregation rules are implemented in a recursive way from

$$Q_t = \Xi_{t-1} Q_{t-1} + \frac{1}{3} y_t^{(3)} \ ,$$

where $\Xi_{t-1} = 0$ in the 1^{st} month and $\Xi_{t-1} = 1$ otherwise (see Harvey, 1989: 309ff). As a result, the required identities hold in the 3^{rd} month of the quarter, with $y_t^Q = Q_t$.

The equation is to be pre-multiplied by the inverse of the left-hand matrix to achieve the standard state space form.

The observation equation is given by

$$\begin{bmatrix} x_t \\ y_t^Q \end{bmatrix} = \begin{bmatrix} \Lambda & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_t \\ y_t \\ y_{t-1} \\ y_t^{(3)} \\ Q_t \end{bmatrix} + \begin{bmatrix} \xi_t \\ 0 \end{bmatrix}$$

The second row, related to y_t^Q, is skipped in months 1 and 2 of the quarter.

Table A.1: Data Euro Area

No.	Series	Publi-cation lag (months)	Trans-formation code	Ranking prediction weights	Ranking LARS
1	Index of notional stock - Money M1	1	2	40	
2	Index of notional stock - Money M2	1	2	37	30
3	Index of notional stock - Money M3	1	2	48	
4	Index of Loans	1	2		24
5	ECB Nominal effective exch. rate	0	2	5	17
6	ECB Real effective exch. rate CPI deflated	0	2	4	15
7	ECB Real effective exch. rate producer prices deflated	0	2		
8	Exch. rate: USD/EUR	0	2	60	
9	Exch. rate: GBP/EUR	0	2		
10	Exch. rate: YEN/EUR	0	2	12	
11	World market prices of raw materials in Euro, total, HWWA	2	2	13	
12	World market prices of raw materials in Euro, total, excl energy, HWWA	2	2	6	13
13	World market prices, crude oil, USD, HWWA	0	2	32	
14	Gold price, USD, fine ounce	0	2	35	19
15	Brent Crude, 1 month fwd, USD/BBL converted in euro	0	2	27	
16	Retail trade, except of motor vehicles and motorcycles	2	2		14
17	IP-Total industry	2	2	20	3
18	IP-Total Industry (excl construction)	2	2	18	2
19	IP-Manufacturing	2	2	17	
20	IP-Construction	2	2	57	
21	IP-Total Industry excl construction and MIG Energy	2	2	54	
22	IP-Energy	2	2		
23	IP-MIG Capital Goods Industry	2	2	59	
24	IP-MIG Durable Consumer Goods Industry	2	2	31	
25	IP-MIG Energy	2	2		
26	IP-MIG Intermediate Goods Industry	2	2	16	
27	IP-MIG Non-durable Consumer Goods Industry	2	2	58	26
28	IP-Manufacture of basic metals	2	2		4
29	IP-Manufacture of chemicals and chemical products	2	2	55	8
30	IP-Manufacture of electrical machinery and apparatus	2	2	21	5
31	IP-Manufacture of machinery and equipment	2	2		
32	IP-Manufacture of pulp, paper and paper products	2	2	28	
33	IP-Manufacture of rubber and plastic products	2	2	19	20
34	Industry Survey: Industrial Confidence Indicator	1	1	3	
35	Industry Survey: Production trend observed in recent months	1	1	14	
36	Industry Survey: Assessment of order-book levels	1	1	2	
37	Industry Survey: Assessment of export order-book levels	1	1	1	
38	Industry Survey: Assessment of stocks of finished products	1	1	10	
39	Industry Survey: Production expectations for the months ahead	1	1	9	
40	Industry Survey: Employment expectations for the months ahead	1	1	11	
41	Industry Survey: Selling price expectations for the months ahead	1	1	15	
42	Consumer Survey: Consumer Confidence Indicator	1	1	24	23
43	Consumer Survey: General economic situation over last 12 months	1	1	22	
44	Consumer Survey: General economic situation over next 12 months	1	1	23	21
45	Consumer Survey: Price trends over last 12 months	1	1	36	11
46	Consumer Survey: Price trends over next 12 months	1	1	53	28
47	Consumer Survey: Unemployment expectations over next 12 months	1	1	25	22
48	Construction Survey: Construction Confidence Indicator	1	1	39	
49	Construction Survey: Trend of activity compared with preceding months	1	1	44	29
50	Construction Survey: Assessment of order books	1	1	43	27
51	Construction Survey: Employment expectations for the months ahead	1	1	38	9
52	Construction Survey: Selling price expectations for the months ahead	1	1	50	1
53	Retail Trade Survey: Retail Confidence Indicator	1	1	47	
54	Retail Trade Survey: Present business situation	1	1	52	10
55	Retail Trade Survey: Assessment of stocks	1	1		
56	Retail Trade Survey: Expected business situation	1	1	49	
57	Retail Trade Survey: Employment expectations	1	1	41	7
58	New passenger car registrations	1	2	33	12
59	Eurostoxx 500	0	2	8	
60	Eurostoxx 325	0	2	7	
61	US S&P 500 composite index	0	2		16
62	US, Dow Jones, industrial average	0	2	56	
63	US, Treasury Bill rate, 3-month	0	1	34	25
64	US Treasury notes & bonds yield, 10 years	0	1	26	
65	Money M2 in the U.S.	1	2		6
66	US, Unemployment rate	1	1	42	
67	US, IP total excl construction	1	2	45	
68	US, Employment, civilian	1	2	46	
69	US, Production expectations in manufacturing	1	1	29	
70	US, Consumer expectations index	0	1	30	
71	10-year government bond yield	0	1	51	18

Transformation code: 1 = monthly difference, 2 = monthly growth rate

Rankings: Ranking of series in stepwise selection (1 = added first / eliminated last)

Table A.2: Data Germany

No.	Series	Publi-cation lag (months)	Trans-formation code	Ranking prediction weights	Ranking LARS
1	Index of notional stock - Money M1	1	2	36	17
2	Index of notional stock - Money M2	1	2	54	
3	Index of notional stock - Money M3	1	2	31	
4	Index of Loans	1	2	58	11
5	ECB Nominal effective exch. rate	1	2		28
6	ECB Real effective exch. rate CPI deflated	1	2	56	
7	ECB Real effective exch. rate producer prices deflated	1	2	55	
8	Exch. rate: USD/EUR	1	2	50	
9	Exch. rate: GBP/EUR	1	2	24	
10	Exch. rate: YEN/EUR	1	2		
11	World market prices of raw materials in Euro, total, HWWA	2	2	32	
12	World market prices of raw materials in Euro, total, excl energy, HWWA	2	2	30	24
13	World market prices, crude oil, USD, HWWA	1	2	42	
14	Gold price, USD, fine ounce	1	2	28	6
15	Brent Crude, 1 month fwd, USD/BBL converted in euro	0	2	34	27
16	IP-Total industry	2	2	16	
17	IP-Total Industry (excl construction)	2	2	20	
18	IP-Manufacturing	2	2	33	
19	IP-Construction	2	2	46	9
20	IP-Total Industry excl construction and MIG Energy	2	2	17	
21	IP-Energy	2	2		
22	IP-MIG Capital Goods Industry	2	2		15
23	IP-MIG Durable Consumer Goods Industry	2	2	40	13
24	IP-MIG Energy	2	2	60	25
25	IP-MIG Intermediate Goods Industry	2	2	38	
26	IP-MIG Non-durable Consumer Goods Industry	2	2	59	4
27	IP-Manufacture of basic metals	2	2	41	2
28	IP-Manufacture of chemicals and chemical products	2	2	47	21
29	IP-Manufacture of electrical machinery and apparatus	2	2	45	12
30	IP-Manufacture of machinery and equipment	2	2		
31	IP-Manufacture of pulp, paper and paper products	2	2		
32	IP-Manufacture of rubber and plastic products	2	2	37	
33	Industry Survey: Industrial Confidence Indicator	1	1	6	
34	Industry Survey: Production trend observed in recent months	1	1	25	
35	Industry Survey: Assessment of order-book levels	1	1	8	
36	Industry Survey: Assessment of export order-book levels	1	1	13	
37	Industry Survey: Assessment of stocks of finished products	1	1	10	26
38	Industry Survey: Production expectations for the months ahead	1	1		18
39	Industry Survey: Employment expectations for the months ahead	1	1	53	3
40	Industry Survey: Selling price expectations for the months ahead	1	1	22	
41	Consumer Survey: Consumer Confidence Indicator	1	1	5	
42	Consumer Survey: General economic situation over last 12 months	1	1	4	
43	Consumer Survey: General economic situation over next 12 months	1	1	3	
44	Consumer Survey: Price trends over last 12 months	1	1	23	
45	Consumer Survey: Price trends over next 12 months	1	1	11	16
46	Consumer Survey: Unemployment expectations over next 12 months	1	1	2	
47	Construction Survey: Construction Confidence Indicator	1	1	7	
48	Construction Survey: Trend of activity compared with preceding months	1	1	15	
49	Construction Survey: Assessment of order books	1	1	1	20
50	Construction Survey: Employment expectations for the months ahead	1	1	9	7
51	Construction Survey: Selling price expectations for the months ahead	1	1	12	1
52	Retail Trade Survey: Retail Confidence Indicator	1	1		
53	Retail Trade Survey: Present business situation	1	1	35	
54	Retail Trade Survey: Assessment of stocks	1	1	43	
55	Retail Trade Survey: Expected business situation	1	1	26	8
56	Retail Trade Survey: Employment expectations	1	1		10
57	New passenger car registrations	1	2	14	22
58	Index of Employment, Construction	3	2		30
59	Index of Employment, Manufacturing	3	2		
60	Eurostoxx 500	0	2	19	14
61	Eurostoxx 325	0	2	18	
62	US S&P 500 composite index	0	2	39	
63	US, Dow Jones, industrial average	1	2	21	
64	US, Treasury Bill rate, 3-month	1	1	57	
65	US Treasury notes & bonds yield, 10 years	1	1	48	
66	Money M2 in the U.S.	1	2	44	19
67	US, Unemployment rate	1	1	51	
68	US, IP total excl construction	1	2		
69	US, Employment, civilian	1	2	49	5
70	US, Production expectations in manufacturing	1	1	27	23
71	US, Consumer expectations index	0	1	52	
72	10-year government bond yield	1	1	29	29

Transformation code: 1 = monthly difference, 2 = monthly growth rate

Rankings: Ranking of series in stepwise selection (1 = added first / eliminated last)

Table A.3: Data France

No.	Series	Publi-cation lag (months)	Trans-formation code	Ranking prediction weights	Ranking LARS
1	Index of notional stock - Money M1	1	2	49	
2	Index of notional stock - Money M2	1	2	45	
3	Index of notional stock - Money M3	1	2	34	
4	Index of Loans	1	2		8
5	ECB Nominal effective exch. rate	1	2	30	
6	ECB Real effective exch. rate CPI deflated	1	2	29	
7	ECB Real effective exch. rate producer prices deflated	1	2	52	
8	Exch. rate: USD/EUR	1	2		16
9	Exch. rate: GBP/EUR	1	2	27	19
10	Exch. rate: YEN/EUR	1	2		22
11	World market prices of raw materials in Euro, total, HWWA	2	2		21
12	World market prices of raw materials in Euro, total, excl energy, HWWA	2	2	48	
13	World market prices, crude oil, USD, HWWA	1	2	59	
14	Gold price, USD, fine ounce	1	2	28	12
15	Brent Crude, 1 month fwd, USD/BBL converted in euro	0	2	32	
16	IP-Total industry	2	2	21	6
17	IP-Total Industry (excl construction)	2	2	22	2
18	IP-Manufacturing	2	2	20	
19	IP-Construction	2	2	41	
20	IP-Total Industry excl construction and MIG Energy	2	2	50	
21	IP-Energy	2	2		11
22	IP-MIG Capital Goods Industry	2	2	51	4
23	IP-MIG Durable Consumer Goods Industry	2	2	55	
24	IP-MIG Energy	2	2		
25	IP-MIG Intermediate Goods Industry	2	2	19	
26	IP-MIG Non-durable Consumer Goods Industry	2	2	44	23
27	IP-Manufacture of basic metals	2	2	43	
28	IP-Manufacture of chemicals and chemical products	2	2	42	
29	IP-Manufacture of electrical machinery and apparatus	2	2	53	26
30	IP-Manufacture of machinery and equipment	2	2	38	17
31	IP-Manufacture of pulp, paper and paper products	2	2	37	
32	IP-Manufacture of rubber and plastic products	2	2	23	7
33	Industry Survey: Industrial Confidence Indicator	1	1	5	
34	Industry Survey: Production trend observed in recent months	1	1	7	
35	Industry Survey: Assessment of order-book levels	1	1	6	
36	Industry Survey: Assessment of export order-book levels	1	1	12	
37	Industry Survey: Assessment of stocks of finished products	1	1	15	28
38	Industry Survey: Production expectations for the months ahead	1	1	14	
39	Industry Survey: Employment expectations for the months ahead	1	1	36	9
40	Industry Survey: Selling price expectations for the months ahead	1	1	16	
41	Consumer Survey: Consumer Confidence Indicator	1	1	9	
42	Consumer Survey: General economic situation over last 12 months	1	1	8	30
43	Consumer Survey: General economic situation over next 12 months	1	1	10	
44	Consumer Survey: Price trends over last 12 months	1	1		14
45	Consumer Survey: Price trends over next 12 months	1	1		20
46	Consumer Survey: Unemployment expectations over next 12 months	1	1	11	
47	Construction Survey: Construction Confidence Indicator	1	1	4	1
48	Construction Survey: Trend of activity compared with preceding months	1	1	3	
49	Construction Survey: Assessment of order books	1	1	2	3
50	Construction Survey: Employment expectations for the months ahead	1	1	1	
51	Construction Survey: Selling price expectations for the months ahead	1	1	13	13
52	Retail Trade Survey: Retail Confidence Indicator	1	1	18	
53	Retail Trade Survey: Present business situation	1	1	17	
54	Retail Trade Survey: Assessment of stocks	1	1		27
55	Retail Trade Survey: Expected business situation	1	1	24	25
56	Retail Trade Survey: Employment expectations	1	1		
57	New passenger car registrations	1	2	54	10
58	Unemployment rate, total	2	1	33	29
59	US, Dow Jones, industrial average	1	2	58	
60	US, Treasury Bill rate, 3-month	1	1	39	
61	US Treasury notes & bonds yield, 10 years	1	1	31	15
62	Money M2 in the U.S.	1	2	56	5
63	US, Unemployment rate	1	1	47	24
64	US, IP total excl construction	1	2	40	
65	US, Employment, civilian	1	2	46	
66	US, Production expectations in manufacturing	1	1		
67	US, Consumer expectations index	0	1	25	18
68	Eurostoxx 500	1	2	60	
69	Eurostoxx 325	1	2	26	
70	US S&P 500 composite index	1	2	57	
71	10-year government bond yield	1	1	35	

Transformation code: 1 = monthly difference, 2 = monthly growth rate

Rankings: Ranking of series in stepwise selection (1 = added first / eliminated last)

Acknowledgements
The present article has originally been published in Hillebrand, E. and S.J. Koopman (eds.), Advances in Econometrics, Vol. 35 by Emerald Insight, ISSN 0731-9053. The author would like to thank Marta Bańbura, Kirstin Hubrich, Christian Schumacher, Bernd Schwaab and two anonymous referees for helpful discussions.

Gerhard Rünstler
European Central Bank; email: gerhard.ruenstler@ecb.europa.eu

ISSN 1725-2806 (online)
ISBN 978-92-899-2022-3
DOI 10.2866/25250
EU catalogue No QB-AR-16-010-EN-N