Rünstler, Gerhard

Working Paper

On the Design of Data Sets for Forecasting with Dynamic Factor Models

WIFO Working Papers, No. 376

Provided in Cooperation with:
Austrian Institute of Economic Research (WIFO), Vienna

Suggested Citation: Rünstler, Gerhard (2010) : On the Design of Data Sets for Forecasting with Dynamic Factor Models, WIFO Working Papers, No. 376, Austrian Institute of Economic Research (WIFO), Vienna

This Version is available at:
https://hdl.handle.net/10419/128924

WIFO
ÖSTERREICHISCHES INSTITUT
FÜR WIRTSCHAFTSFORSCHUNG

WORKING PAPERS

On the Design of Data Sets for Forecasting with Dynamic Factor Models

Gerhard Rünstler

376/2010

On the Design of Data Sets for Forecasting with Dynamic Factor Models

Gerhard Rünstler

WIFO Working Papers, No. 376
July 2010

E-mail address: Gerhard.Ruenstler@wifo.ac.at
2010/212/W/0

Medieninhaber (Verleger), Hersteller: Österreichisches Institut für Wirtschaftsforschung • 1030 Wien, Arsenal, Objekt 20 • Tel. (43 1) 798 26 01-0 • Fax (43 1) 798 93 86 • http://www.wifo.ac.at/ • Verlags- und Herstellungsort: Wien

Kostenloser Download:
http://www.wifo.ac.at/wwa/jsp/index.jsp?fid=23923&id=40093&typeid=8&display_mode=2

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

Gerhard Rünstler*

Austrian Institute for Economic Research

July 2010

Abstract

Forecasts from dynamic factor models potentially benefit from refining the data set by eliminating uninformative series. The paper proposes to use forecast weights as provided by the factor model itself for this purpose. Monte Carlo simulations and an empirical application to forecasting euro area, German, and French GDP growth from unbalanced monthly data suggest that both forecast weights and Least Angle Regressions result in improved forecasts. Overall, forecast weights provide yet more robust results.

Keywords: Dynamic factor models, forecasting, variable selection, LARS

JEL classification: E37, C53, C51

.

*Austrian Institute for Economic Research, Arsenal Objekt 20, A-1030 Vienna, Gerhard.Ruenstler@wifo.ac.at

1 Introduction

Dynamic factor models have emerged as a widely used tool for obtaining short-term forecasts of economic activity and inflation (e.g. Stock and Watson, 2002; Banerjee et al., 2005; 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. Recently, 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 worsen the precision of factor estimates and have proposed heuristic rules to eliminate redundant series. Bai and Ng (2008) have advocated Least Angle Regressions (LARS) and related methods for identifying efficient sets of predictors in dynamic factor models. These two studies, along with Caggiano et al. (2009) and Schumacher (2010), also present empirical applications, which demonstrate that using smaller data sets may improve forecasts from dynamic factor models.

In this paper, I investigate an alternative method for selecting an efficient set of predictors in a dynamic factor model. I propose the use of forecast weights, which are obtained from the factor model itself. As with any linear model, the factor model forecast for a certain target variable can be written as a weighted linear combination of current and lagged values of the predictors. I will investigate, whether forecast efficiency can be improved by retaining only predictors with high forecast 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 predictors 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 (see Efron et al., 2004). Bai and Ng (2008) report considerable gains in forecast performance from using these methods to select predictors in factor model forecasts of inflation.

Another way to deal with high dimensionality is to use the factor model itself for approximating

the marginal predictive gains of individual series. This can be achieved from calculating the weights of the individual series in the factor model forecast.

I will provide two pieces of evidence, which suggest that these weights are a useful alternative to LARS. First, a Monte-Carlo simulation exercise confirms that, under certain conditions, both methods are suitable for selecting data sets that result in more efficient forecasts. However, forecast weights are more successful than LARS in identifying the appropriate series. Consequently, they also tend to deliver better out-of-sample forecasts. LARS, in turn, appears to suffer from overfitting, as in-sample forecasts suggest gains in forecast precision that only partly carry over to the out-of-sample forecasts and may result in misleading choices on efficient sets of predictors.

Second, I will apply both methods in forecasting quarterly GDP growth from large unbalanced monthly data sets. I use the dynamic factor model by Doz et al. (2005), which differs from other versions of dynamic factor models in that factor dynamics is explicitly modeled and the Kalman smoother is used to obtain forecasts. As a result, the model 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; Angelini et al., 2008; Rünstler et al., 2009). Forecast weights of individual series can be obtained from the Kalman smoother (Bańbura and Rünstler, 2010). LARS is less suited for dealing with such conditions and must be applied to quarterly aggregates of monthly data, while ignoring unbalancedness.

I use unbalanced monthly data sets for the euro area, Germany, and France over the period of 1991 to 2007. Each data set contains about 70 series. Model specifications and selections are obtained from a pre-sample. I find that variable selections from either method improve forecast performance for the euro area and Germany with small data sets of about 10 to 20 series, but not so for France. Again, forecast weights provide more robust variable selections and give smaller out-of-sample forecast errors than LARS.

The paper is organised as follows. Section 2 discusses the use of forecast weights in a static context and conducts the Monte Carlo study to investigate the performance of forecast weights and LARS. Section 3 presents the empirical application. Section 4 concludes the paper.

2 Monte Carlo simulation

This section conducts a Monte Carlo study to investigate the gains from variable selection methods in forecasting with a static factor model

$$\begin{aligned} x_t &= \lambda f_t + \xi_t, \qquad t = 1, \ldots, T \\ f_t &\sim \mathbb{N}(0, 1) \\ \xi_t &\sim \mathbb{N}(0, \Sigma_\xi) \end{aligned}$$

where the $n \times 1$ vector of series $x_t = (x_{1,t}, \ldots, x_{n,t})'$ is related to a single latent factor f_t by a vector $\lambda = (\lambda_1, \ldots, \lambda_n)'$ of factor loadings. Both factor f_t and idiosyncratic components ξ_t are white noise. The purpose is to forecast series y_t from the latent factor,

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

where $\varepsilon_t \sim \mathbb{N}(0, \sigma_\varepsilon)$ and $\sigma_\varepsilon^2 = 1 - \beta^2$, which implies $\text{var}(y_t) = 1$.

Asymptotic theory suggests that the factor space can be consistently estimated by principal components as $[n; T] \to \infty$ if (i) the errors are stationary, (ii) the factors have non-trivial loadings, and (iii) the idiosyncratic errors have weak correlation both serially and cross-sectionally (Stock and Watson, 2002; Bai and Ng, 2002). Specifically, under the third condition, the non-diagonal elements of Σ_ξ should converge towards zero, as n tends to infinity, and the diagonal terms should approach the cross-section idiosyncratic variance.

Boivin and Ng (2006) have argued that the third condition is likely to be violated in macroeconomic data sets. Moreover, they have demonstrated that in case of non-zero cross-correlations and heteroscedasticity in idiosyncratic components, the precision of factor estimates does not necessarily increase with the number of series. One may therefore improve factor estimates by eliminating series with idiosyncratic components that are subject to high variance and high mutual cross correlations.

The below simulation inspects the gains in forecast precision from two variable selection methods under these conditions. I will use forecast weights as obtained from the factor model and Least

Angle Regressions (LARS). The simulation design is a variant of simulation 1 in Boivin and Ng (2006). I define two groups of series in x_t, which differ by their factor loadings, i.e.

$$\lambda_i = \begin{cases} \lambda^A, & \text{for } i = 1, \ldots, m \\ \lambda^B < \lambda^A, & \text{for } i = m+1, \ldots, n \end{cases}$$

This implies heteroscedasticity in the respective idiosyncratic components, as series x_{it} are standardised to $\text{var}(x_{i,t}) = 1$. 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 $i, j = 1, \ldots, n, \quad i \neq j$. Hence,

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

For non-zero values of ρ forecast precision may be improved from using reduced data sets. In the present simulation design, clearly series $x_{1,t}$ to $x_{m,t}$ are more informative for factor estimates than the remaining series. Hence, any variable selection method should prefer those series over $x_{i,t}$, $i > m$.

I turn to variable selection methods. In a static context, with factors being estimated by principal components, calculating the forecast weights $\widehat{\omega}$ implied by the model is straightforward. Let $(1/T)\sum_{t=1}^{T} x_t x_t' = VDV'$ be the eigenvalue decomposition of the empirical correlation matrix with eigenvectors V. Given a certain number r of factors, it holds $\widehat{\lambda} = V_r$ and $\widehat{f}_t = V_p' x_t$, where V_r denotes the first r colums of V. The forecast for y_t is then found with

$$\widehat{y}_t = \widehat{\beta} V_p' x_t = \widehat{\omega}' x_t$$

where $\widehat{\beta}$ is estimated by OLS from y_t on $\widehat{f}_t$.

For a factor model, stepwise backward elimination seems a natural approach. I start with the entire set of series $x_t^{W,n} = x_t$. At each step $i = n, \ldots, 1$, I re-estimate the factor model based on series $x_t^{W,i}$. I then calculate forecast weights and remove the series with the lowest weight from $x_t^{W,i}$ to obtain selection $x_t^{W,i-1}$. Overall, I obtain a set of selections $\mathcal{W} = \left\{x_t^{W,i}\right\}_{i=1}^{n}$. The number of factors is kept fixed in the selection process.

LARS is a constrained variant of stepwise forward selection, which is more efficient by using a conservative choice of the coefficients in the forecast equation (Efron et al., 2004). Starting with

empty set $x_t^{L,0}$, at each step i one series is added to $x_t^{L,i-1}$ to obtain $x_t^{L,i}$. As with stepwise forward selection, this is the series that displays the highest correlation with the residual from the forecast equation using set $x_t^{L,i-1}$. In LARS, at each step after adding a series, the coefficients in the forecast equation based on $x_t^{L,i}$ are adjusted in a certain way to improve the robustness of variable selection. This is achieved by increasing the coefficients in their joint least squares direction until another predictor (not yet contained in $x_t^{L,i}$) displays as much correlation with the residual as the series in $x_t^{L,i}$. The process stops at $k = \min(T, n-1)$, which results in a set of selections $\mathcal{L} = \left\{x_t^{L,i}\right\}_{i=1}^{k}$.

Bai and Ng (2008) have used LARS and some of its variants (i.e. LASSO and elastic net algorithms, see Efron et al., 2004) to select series in factor model forecasts for U.S. inflation. They report that LARS performs well over a range of specifications.

I use 1000 draws of $n = 100$ series. For each draw $\{x_{t,J}, y_{t,J}\}$ I proceed as follows:

1. I split the draw of length T into two subsamples 1 and 2 of length T_1 and T_2, respectively, with $T_1 + T_2 = T$.
2. From sub-sample 1, I obtain the sets of selections $\mathcal{W}_J = \left\{x_{t,J}^{W,i}\right\}_{i=1}^{n}$ and $\mathcal{L}_J = \left\{x_{t,J}^{L,i}\right\}_{i=1}^{k}$ according to forecast weights and LARS, respectively. All model parameters are estimated from sub-sample 1. The number of factors r is estimated from the full data set using information criterion PC_{p2} from Bai and Ng (2002).[1]
3. In-sample and out-of-sample forecasts for y_t^J are obtained from subsamples 1 and 2, respectively. Specifically, I obtain forecasts $\widehat{y}_{t,J}^{W,i}$ and $\widehat{y}_{t,J}^{L,i}$ for y_t^J over both subsamples, based on selections contained in $\mathcal{W}_J$ and $\mathcal{L}_J$. Again, all parameters are taken from sub-sample 1.

I set $T = 180$, which amounts to 15 years of monthly data, while $T_1 = T_2 = 90$. Factor loadings are set to $\lambda^A = \beta = 0.7$ and $\lambda^B = 0.3$. I vary the share m of series with high loading λ^A in between 0.1 and 0.5. Correlation ρ is set to values between 0.0 and 0.3.

Table 1 shows the findings from the Monte Carlo exercise for in-sample and out-of-sample forecasts. I present two sets of results. The first set, shown in the upper panel of the table, is based

[1] For LARS I use MATLAB code provided by Karl Skoglund (www.cs.uiuc.edu/~dengcai2/Data/code/lars.m).

on an in-sample miminum RMSE criterion to choose the optimal size of the selections. For each draw, I determine the selections in $\mathcal{W}_J$ and $\mathcal{L}_J$ that give the minimum root mean squared error (RMSE) of forecasts $\widehat{y}_{t,J}^{W,i}$ and $\widehat{y}_{t,J}^{L,i}$ in sub-sample 1. I then apply these selections to the out-of-sample forecasts. The upper panel of Table 1 reports the average RMSE of these forecasts, the number of series chosen by the in-sample RMSE criterion, and the percentage of series of type A (i.e. with loading λ^A) contained in the selections. The second set of results, shown in the lower panel of the table, uses a grid of fixed number of series in the various selections.

The results indicate that selections from forecast weights provide slightly better out-of-sample forecasts. In addition, they are less prone to overfitting, whereas in-sample forecasts from LARS selections suggest spurious gains in forecast precision.

Both selection methods result in improved out-of-sample forecast performance compared to the full data set, but with two exceptions. First, for $\rho = 0$, variable selection has negligible effects, as expected. Second, while the information criterion PC_{p2} chooses the number of factors generally correctly with $r = 1$, on some occasions $r = 2$ is selected for simulations with $\rho = 0.3$ and $m/n \geq 0.3$. In this case, the application of either variable selection method worsens out-of-sample forecast performance. There emerges however a difference between the two methods in whether these losses are properly detected from in-sample forecasts. For forecast weights, outcomes from in-sample and out-of-sample forecasts are similar. By contrast, in-sample forecasts from small LARS selections suggest spurious gains in forecast precision, and hence give misleading signals on the appropriate selections for out-of-sample forecasts.

In the remaining cases both selection methods improve out-of-sample forecast performance. For $m/n = 0.1$ the RMSE declines by about 10%. Forecast weights generally fare somewhat better than LARS. This appears to be related to two differences in the simulation outcomes. First, forecast weights are considerably better in detecting the series with high factor loadings λ^A, as evidenced from the share of these series in the selections. This holds especially for the case of $m/n = 0.1$. Second, the in-sample RMSE criterion chooses LARS selections with rather small number of series even for large values of m/n. These selections deliver in-sample forecasts, which grossly over-estimate the gains that would arise in out-of-sample forecasts. By contrast, for

forecast weights the selected number of series tends to be close to m, while in-sample forecast gains reflect quite closely those out of sample.

3 Forecasting GDP growth from monthly data

This section presents a pseudo real-time exercise to forecast GDP growth of the euro area, Germany, and France from large monthly data sets, each containing about 70 series over the period of 1991 to 2007. The forecast design takes account of the different timing of data releases of the individual series, which results in unbalanced data availability at the end of the sample.

I use the dynamic factor model by Doz et al. (2005), which models factor dynamics explicitly and uses a state-space representation to estimate and forecast the latent factors from the Kalman smoother. As a by-product, the model handles mixed frequencies and unbalanced data sets in an efficient way. It has been shown to perform well compared to other methods in forecasting GDP growth of the euro area and several member states (Angelini et al., 2008; Rünstler et al., 2009). These studies report that differences in the timing of data releases among individual series have large effects on their marginal predictive gains within a certain data set.

I will use both forecast weights and LARS to obtain variable selections from the first half of the sample and investigate the forecast performance in the second part of the sample.

3.1 Factor model

The model by Doz et al. (2005) is given by the equations

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

$$f_{t+1} = \sum_{s=1}^{p} A_i f_{t-s+1} + B\eta_t, \qquad \eta_t \sim \mathbb{N}(0, I_q), \tag{2}$$

where $x_t = (x_{1t}, \ldots, x_{nt})'$, $t = 1, \ldots, T$, is a vector of n stationary monthly series, which have been standardised to mean zero and variance one.

Equation (1) relates the monthly series x_t to a $r \times 1$ vector of latent factors $f_t = (f_{1,t}, \ldots, f_{r,t})'$

from a matrix of factor loadings Λ plus an idiosyncratic component $\xi_t = (\xi_{1,t}, \ldots, \xi_{n,t})'$. The latter is assumed to be multivariate white noise with covariance matrix Σ_ξ. Equation (2) describes the law of motion for the latent factors f_t, which are driven by q-dimensional white noise η_t, where B is a $r \times q$ matrix. The stochastic process for f_t is assumed to be stationary.[2]

The model can be cast in state space form. Estimates and forecasts of factors f_t are obtained from the Kalman smoother. One advantage of this approach is that the Kalman smoother efficiently handles mixed frequency data. I use the model implementation by Angelini et al. (2010), which introduces monthly GDP growth y_t as a latent variable in the state space form. Data x_t represent monthly rates of change. Consequently, y_t is assumed to be related to factors f_t by equation

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

This is supplemented with log-linear aggregation rules to relate y_t to observed quarterly GDP growth. Define series y_t^Q at monthly frequency such that it contains observed quarterly GDP growth y_{3k}^Q in the 3^{rd} month of the respective quarter, whereas the remaining observations are treated as missing. Aggregation rules can then be expressed as

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

where $y_t^{(3)}$ denotes 3-month growth rates of monthly GDP growth, i.e. the growth rates vis-a-vis the same month of the previous quarter. The state space form is given in the annex.

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 as principal components of the empirical correlation matrix of x_t. The latter are used to estimate A_i in equation (2), while a further application of principal components to the residual covariance matrix of the VAR gives matrix B. Parameters β and σ_ε^2 are estimated from a quarterly version of equation (3), again using

[2] The representation given by equation (1) is static in the sense that x_t load only on current values of factors. However, this representation can be derived from a restricted version of a general DFM with q dynamic factors where x_t loads current and lagged values (see Stock and Watson, 1995).

the initial estimates of factors f_t [3] I 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 AIC, and criterion 2 in Bai and Ng (2007), respectively.

Another advantage of the Kalman smoother is that it efficiently handles unbalanced data sets. Let $z_t' = (x_t', y_t^Q)$ and denote with $\mathcal{Z}_t$ the information set in period t. Real-time data sets typically contain missing observations at the end of the sample due to publication lags. In the forecast exercise I will follow former studies (e.g. Giannone et al., 2008) and apply *pseudo-real time* data sets $\mathcal{Z}_t$, which use the final data releases but take account of the timing of data releases. 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 $x_{i,t-s}$, $s \geq 0$, if and only if observation $x_{i,T-s}$ is missing in $\mathcal{Z}_T$, $i = 1, \ldots, n$. Quarterly GDP growth is treated in an equivalent way.

For state-space form

$$\begin{aligned} z_t &= W_t \alpha_t + u_t \qquad u_t \sim \mathbb{N}(0, \Sigma_u) \\ \alpha_{t+1} &= T_t \alpha_t + v_t, \qquad v_t \sim \mathbb{N}(0, \Sigma_v) \end{aligned} \tag{4}$$

the Kalman smoother provides minimum mean square linear (MMSE) estimates $a_{t+h|t} = \mathbb{E}\left[\alpha_{t+h} | \mathcal{Z}_t\right]$ of the state vector and their covariance $P_{t+h|h}$ for any $h > -t$. To handle missing observations, the rows in equation (4) corresponding to missing observations in z_t are simply skipped when applying the Kalman smoother recursions (Durbin and Koopman, 2003:92f).

Bańbura and Rünstler (2010) have applied an algorithm by Harvey and Koopman (2003) to obtain the Kalman smoother weights of individual observations in estimates $a_{t+h|t}$. For forecasts based on pseudo real-time data sets $\mathcal{Z}_t$, these weights become independent of time t, as the Kalman filter approaches its steady state. Since y_t^Q is an element of the state vector, this allows expressing estimates and forecasts $y_{t+h|t}^Q$ as

$$y_{t+h|t}^Q = \sum_{s=0}^{t-1} \omega_s'(h) z_{t-s} \ . \tag{5}$$

[3] Estimation requires a complete data set und is therefore done from a balanced sample. Banbura and Modugno (2009) present an EM algorithm to estimate the model in case of missing data.

with weights $\omega_s(h)$. Since the Kalman smoother provides MMSE estimates, weights $\omega_{i,s}(h)$ are the measure of the marginal predictive gain in $y^Q_{t+h|t}$ that arises from adding series $x_{i,t-s}$ to the data set. In the below forecast exercise, I will use cumulative weights $\omega(h) = \sum_{s=0}^{t-1} \omega_s(h)$.[4]

3.2 Forecast design

I use monthly data sets for the euro area, Germany, and France of each about 70 series. All data sets start in January 1991 and were downloaded on 15, Dec 2009. The choice of series is based on Angelini et al. (2008) and includes national data on economic activity (such as industrial production, trade, employment), the European Commission business and consumer surveys and data on the international economy including financial markets. European Commission surveys and financial market data are available right at the end of the respective month. By contrast, most of the official data on economic activity are published with a delay of 6 to 8 weeks after the end of the month. The same applies to the euro area monetary aggregates.[5]

All series have been transformed to monthly rates of change and partly taken in logs beforehand. Further, the data have been cleaned from outliers. The series are listed in annex A together with their publication lags and the data transformations used.

The forecast design aims at replicating the real-time application of the factor model as closely as possible. I account for the timing of data releases of the individual the series. by using data sets $\mathcal{Z}_t$, as defined above. I inspect *seven* forecasts for GDP growth in in a certain quarter, which are obtained in consecutive months. I start with forecasting in the 1^{st} month of the previous quarter and stop in the 1^{st} month of the subsequent (next) quarter, one month before the flash estimate of GDP is released. To forecast GDP growth in the 2^{nd} quarter, for instance, the 1^{st} forecast is run in January and the final (7^{th}) one in August. Note that 'forecast' 7 is actually a backcast, whereas 'forecasts' 4 to 6 amount to nowcasting the current quarter. Forecasts 1 to 3 amont to

[4] Cumulative weights do not precisely measure the predictive gain of a series across all lags. 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 smoother.

[5] I am grateful to M. Bańbura for providing me with the data.

forecasting GDP growth in the next quarter.

I obtain the variable selections and corresponding factor model specifications from a first part of the sample and run a recursive forecast exercise on the remainder. I proceed as follows:

1. I obtain selections $\left\{x_t^{W,i}\right\}_{i=1}^{n}$ and $\left\{x_t^{L,i}\right\}_{i=1}^{k}$ from the data samples ranging until 2000 Q4 using stepwise elimination as described in section 2.1. Selections are based on the Kalman smoother forecast weights from the mid-quarter nowcast (forecast 5) as from pseudo real-time data sets $\mathcal{Z}_t$. To obtain selections from LARS, I aggregate the monthly data to quarterly frequency and apply LARS to static regressions of quarterly GDP growth on the quarterly rates of change of the data.

   For all selections, model specifications are obtained from the information criteria set out in section 3.1. 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 forecast weights I obtain specifications $\left(r^{W,i}, p^{W,i}, q^{W,i}\right)$ related to selection $x_t^{W,i}$ under the restrictions $r^{W,i} \leq r^{W,i+1}$, $p^{W,i} \leq p^{W,i+1}$, and $q^{W,i} \leq q^{W,i+1}$ for $i < n$.

2. I obtain forecasts of GDP growth over the period from 2001 Q2 to 2007 Q4 (27 observations) based on pseudo real-time data sets $\mathcal{Z}_t$ and recursive parameter estimates. I obtain these forecasts for a grid of different variable selections as from above.[6]

3.3 Results

Table 2 shows the RMSE over the period of 2001 Q2 to 2007 Q4 for the sequence of 7 forecasts. The RMSE is shown relative to the naive forecast, which is based on a random walk with drift for GDP. As the naive forecast is based on quarterly data, the RMSE shifts in 3-month terms. The timing of the shifts reflects the publication dates of the GDP flash estimates.

For all three data sets, the factor model forecasts from the entire set of series improve upon the naive forecast, but the gains are moderate for Germany. For the euro area, the results are somewhat worse compared to earlier studies, that use a very similar set of series, but a different

[6] The years 2008 and 2009 have been omitted from the sample due to exceptionally large forecast errors.

sample (e.g. Angelini et al., 2008). For Germany, the results are in line with various studies that report difficulties with obtaining informative forecasts (Marcellino and Schumacher, 2008; Schumacher, 2010).

The specifications chosen by the information criteria in the selection sample are similar across data sets. For the complete data sets, the number of factors is estimated with $r = 3$, while estimates of p and q range between 2 and 3. Once the number of series declines for the selections, estimates of r remain unchanged at 3, while estimates of p and q shrink towards 1. For all countries, the average cross correlation among idiosyncratic components is slightly below 0.2. Idiosyncratic components are subject to considerable heteroscedasticity.

Table 2 shows the RMSE of out-of-sample forecasts based on a grid of variable selections as from forecast weights and LARS. Figure 1 shows the RMSE of the optimal selections. Note that variable selection with LARS stops at $k = \min(T_1, n - 1)$.

Both variable selection methods deliver improved forecasts for the euro area and Germany at the shorter horizons, i.e. forecasts 4 to 7. At the longer horizons (forecasts 1 to 3) all selections perform about equally well compared to the full data sets. For forecasts 4 to 7, the best results arise from small data sets containing about 20 series. For the euro area, the best-performing selections from either method reduce the RMSE by as much as 20%. For Germany, selections from forecast weights with 20 series or less result in gains of close to 15%. LARS selections perform somewhat worse in this case. The best-performing selection with 30 series gives rise to a gain of slightly less than 10%.

No gains from variable selection are found for France. In this case, the smaller data sets with less than 40 series actually perform considerably worse than the full data set at the short horizons.

The question remains whether the appropriate data sets would have been chosen in real time. Table 3 shows the in-sample RMSE for the selection sample (up to 2000 Q4) for the various selections. The in-sample RMSE shows rather little variation across selections and the results are therefore only indicative. Nevertheless, the criterion appears to work reasonably well, but with the exception of LARS selection for France. For the euro area and Germany, in all cases small

data sets of in between 10 and 20 series are selected, which comes close to the optimal choices.

For France, the in-sample criterion would have resulted in the correct choice of 50 series or more for forecast weights. However, the in-sample forecasts from LARS suggest considerable efficiency gains from small data sets, which are not present in the out-of-sample forecasts. The choice of 15 or 20 series would have resulted in a considerable loss in out-of-sample forecast efficiency. This result parallels the findings from the Monte Carlo exercise on the overfitting tendency of LARS.

The ranking of the series according to the selections are shown in Tables A.1 to A.3 in the annex. Selections from forecast weights are less heterogenous than those of LARS. For the euro area, forecast weights select the main items of business (confidence indicators and order books) and consumer surveys, together with equity price indices and the euro area real effective exchange rate. For Germany, survey items are even more prominent. LARS puts stronger emphasis on hard data, such as items of industrial production, retail sales and new passenger car registrations. These differences are to some extent a consequence of the neglect of publication lags in those series by LARS.

Tables 4 and 5 finally present the outcomes of several robustness checks. Table 4 shows the outcomes of the exercise under fixed model specifications. That is, specification (r, p, q) as obtained for the full data set is kept constant for all variable selections. This has very little effects on the results, which might partly arise from the fact that specifications for the entire data set are already rather small.

Table 5 shows next-quarter forecasts from alternative variable selections. The above results have used selections based on nowcasts, i.e. forecast 5 for forecast weights and LARS regressions of quarterly GDP growth on the contemporaneous values of the predictors. However, optimal selections may depend on the forecast horizon h. This would suggest the use of horizon-specific variable selections. Table 5 shows the outcome of an exercise, where variable selections are based on one-quarter ahead forecasts. I use forecast weights for forecast 2 and LARS regressions, where the predictors are lagged by one quarter. As shown in Table 5, however, the alternative weights do not improve forecast performance for either selection method.

4 Conclusions

The paper has inspected the efficiency gains from variable selection when forecasting with a dynamic factor model. I have considered short-term forecasts for GDP growth of the euro area, Germany, and France from unbalanced monthly data sets. I have compared two methods for this purpose, i.e. Least Angle Regressions (LARS), as proposed by Boivin and Ng (2006), and factor model forecast weights.

Against the earlier studies by Bai and Ng (2008), Caggianio et al. (2009) and Schumacher (2010), this paper was the first one to inspect the success of variable selection from a pre-sample - as opposed to in-sample selection - and with unbalanced data sets. The results still confirm earlier findings that variable selection methods tend to improve forecast efficiency. However, this should not be taken for granted in all applications. In the examples investigated in this paper, efficiency gains occured for nowcasts in two of the three data sets. For one-quarter ahead forecasts, variable selection did not systematically affect forecast performance.

Forecast weights performed somewhat better than LARS in an out-of-sample context. Both a Monte Carlo simulation and the empirical application suggest that they provide more robust selections and smaller forecast errors. LARS tends to overfitting. In-sample forecasts suggest gains in forecast precision, which do not necessarily carry over to the out-of-sample forecasts. Hence, some caution is required in application. As a further advantage against LARS, factor model forecast weights are applicable in a wider range of circumstances, such as dynamic models, mixed frequencies and unbalanced data sets.

One question for future research is whether factor models are useful for the pre-screening of variables also in the context of other forecasting methods.

Table 1: Monte Carlo Simulation

Selection size from in-sample RMSE criterion

Correlation (ρ)		0,0	0,0	0,0	0,1	0,1	0,1	0,3	0,3	0,3	0,3	0,3	0,3
m/n		0,1	0,3	0,5	0,1	0,3	0,5	0,1	0,3	0,5	0,1	0,3	0,5
Nr of factors (p)		1	1	1	1	1	1	1	1	1	2	2	2
Nr of series selected													
Weights		52,4	47,4	50,6	9,9	23,1	33,7	10,4	19,6	28,6	22,5	60,8	71,2
LARS		21,5	19,8	18,7	10,5	11,1	12,0	6,2	7,1	7,8	15,7	20,1	21,3
RMSE out-of-sample													
All series		0,75	0,74	0,73	0,86	0,81	0,79	0,93	0,89	0,86	0,84	0,76	0,75
Forecast weights		0,75	0,74	0,74	0,79	0,77	0,77	0,83	0,81	0,81	0,82	0,77	0,76
LARS		0,77	0,76	0,75	0,82	0,79	0,79	0,84	0,83	0,82	0,82	0,81	0,80
RMSE in-sample													
All series		0,73	0,72	0,72	0,84	0,79	0,77	0,91	0,88	0,84	0,78	0,73	0,73
Forecast weights		0,72	0,71	0,71	0,76	0,74	0,74	0,80	0,79	0,82	0,77	0,73	0,73
LARS		0,62	0,62	0,62	0,70	0,68	0,67	0,78	0,76	0,75	0,71	0,67	0,66
% series of type A													
Forecast weights		0,38	0,71	0,86	0,91	0,98	0,99	0,88	0,99	1,00	0,62	0,52	0,62
LARS		0,27	0,52	0,68	0,53	0,79	0,87	0,79	0,94	0,98	0,41	0,52	0,59

Fixed selection sizes

Correlation (ρ)		0,0	0,0	0,0	0,1	0,1	0,1	0,3	0,3	0,3	0,3	0,3	0,3
m/n		0,1	0,3	0,5	0,1	0,3	0,5	0,1	0,3	0,5	0,1	0,3	0,5
Nr of factors (p)		1	1	1	1	1	1	1	1	1	2	2	2
RMSE out-of-sample													
All series	100	0,75	0,74	0,73	0,86	0,81	0,79	0,93	0,89	0,86	0,93	0,76	0,75
Forecast weights	60	0,76	0,74	0,73	0,85	0,79	0,77	0,92	0,86	0,82	0,84	0,76	0,76
	40	0,76	0,74	0,73	0,84	0,78	0,76	0,91	0,83	0,81	0,84	0,77	0,77
	20	0,76	0,74	0,74	0,81	0,77	0,77	0,88	0,81	0,81	0,82	0,79	0,79
	10	0,76	0,76	0,76	0,79	0,78	0,79	0,84	0,82	0,82	0,82	0,81	0,81
LARS	60	0,75	0,74	0,74	0,85	0,82	0,79	0,93	0,89	0,86	0,85	0,77	0,76
	40	0,76	0,75	0,74	0,85	0,81	0,79	0,92	0,89	0,87	0,84	0,78	0,77
	20	0,77	0,76	0,75	0,83	0,80	0,79	0,90	0,86	0,85	0,83	0,81	0,80
	10	0,78	0,76	0,76	0,82	0,79	0,78	0,87	0,84	0,83	0,84	0,83	0,82
RMSE in-sample													
All series	100	0,73	0,72	0,72	0,84	0,79	0,77	0,91	0,88	0,84	0,78	0,73	0,73
Forecast weights	60	0,74	0,72	0,72	0,83	0,78	0,75	0,90	0,84	0,80	0,78	0,74	0,74
	40	0,74	0,72	0,72	0,82	0,76	0,75	0,89	0,81	0,79	0,78	0,75	0,75
	20	0,74	0,73	0,73	0,79	0,76	0,76	0,85	0,80	0,80	0,78	0,77	0,77
	10	0,74	0,74	0,74	0,77	0,77	0,77	0,80	0,80	0,80	0,79	0,73	0,78
LARS	60	0,71	0,71	0,71	0,83	0,79	0,77	0,91	0,88	0,85	0,77	0,72	0,71
	40	0,69	0,69	0,69	0,82	0,78	0,76	0,90	0,86	0,84	0,75	0,69	0,68
	20	0,68	0,67	0,67	0,78	0,74	0,72	0,88	0,83	0,81	0,73	0,68	0,67
	10	0,71	0,69	0,68	0,75	0,72	0,71	0,84	0,78	0,76	0,74	0,74	0,74

In the upper panel of Table 1, the number of series in either selection is determined from an in-sample RMSE criterion (see main text). The average number of series chosen is shown in the upper rows together with the average percentage of series of type A that are contained in the selections. The lower panel shows the results for variable selections of fixed sizes.

Table 2: Out-of sample RMSE

(2001 Q 2 - 2007 Q4)

Euro area

	Naive	AR(1)	Kalman filter weights								LARS			
			All	60	50	40	30	20	15	10	30	20	15	10
7	0,284	0,84	0,84	0,79	0,71	0,71	0,69	0,65	0,70	0,73	0,57	0,56	0,58	0,61
6	0,284	0,84	0,91	0,87	0,79	0,78	0,73	0,69	0,74	0,78	0,66	0,71	0,67	0,71
5	0,284	0,84	0,93	0,92	0,87	0,85	0,80	0,76	0,78	0,82	0,79	0,90	0,90	0,96
4	0,288	0,95	0,83	0,84	0,82	0,80	0,79	0,71	0,74	0,82	0,80	0,95	1,00	1,00
3	0,288	0,95	0,74	0,76	0,75	0,74	0,75	0,65	0,68	0,79	0,78	1,02	1,03	1,04
2	0,288	0,95	0,79	0,79	0,79	0,78	0,78	0,72	0,74	0,86	0,94	1,01	1,01	1,00
1	0,290	0,97	0,77	0,78	0,79	0,77	0,77	0,72	0,76	0,88	0,94	0,99	0,99	0,99
4-7	0,285	0,87	0,88	0,86	0,80	0,79	0,75	0,70	0,74	0,79	0,71	0,78	0,79	0,82
1-3	0,288	0,96	0,77	0,78	0,77	0,76	0,77	0,69	0,72	0,84	0,89	1,01	1,01	1,01
1-7	0,286	0,91	0,83	0,82	0,79	0,78	0,76	0,70	0,73	0,81	0,78	0,88	0,88	0,90

Germany

	Naive	AR(1)	Kalman filter weights								LARS			
			All	60	50	40	30	20	15	10	30	20	15	10
7	0,475	1,00	1,01	1,05	1,04	0,99	0,93	0,80	0,80	0,80	0,76	0,80	0,81	0,88
6	0,475	1,00	0,97	1,00	0,99	0,96	0,93	0,78	0,78	0,79	0,80	0,84	0,86	0,89
5	0,475	1,00	0,95	0,97	0,96	0,92	0,89	0,81	0,81	0,83	0,88	0,92	0,95	0,90
4	0,479	1,01	0,90	0,91	0,91	0,93	0,93	0,89	0,90	0,91	0,92	0,97	0,97	0,97
3	0,479	1,01	0,85	0,86	0,86	0,87	0,86	0,88	0,87	0,86	0,97	0,98	0,98	0,97
2	0,479	1,01	0,89	0,89	0,90	0,91	0,90	0,91	0,91	0,92	0,98	0,98	0,98	0,98
1	0,481	1,01	0,87	0,88	0,88	0,91	0,91	0,91	0,95	0,96	0,97	0,97	0,97	0,98
4-7	0,476	1,00	0,95	0,98	0,98	0,95	0,92	0,82	0,82	0,83	0,84	0,88	0,90	0,91
1-3	0,480	1,01	0,87	0,88	0,88	0,90	0,89	0,90	0,91	0,91	0,97	0,98	0,98	0,98
1-7	0,478	1,01	0,92	0,94	0,94	0,93	0,91	0,85	0,86	0,87	0,90	0,92	0,93	0,94

France

	Naive	AR(1)	Kalman filter weights								LARS			
			All	60	50	40	30	20	15	10	30	20	15	10
7	0,338	1,16	0,65	0,67	0,67	0,84	0,86	0,91	0,93	1,15	0,69	0,75	0,79	0,76
6	0,338	1,16	0,69	0,70	0,70	0,85	0,86	0,89	0,93	1,22	0,87	0,93	0,90	0,93
5	0,338	1,16	0,80	0,77	0,75	0,90	0,92	0,87	1,00	1,32	1,12	1,18	1,15	1,21
4	0,338	1,01	0,88	0,89	0,91	0,85	0,84	0,77	0,90	1,18	1,11	1,12	1,21	1,25
3	0,338	1,01	0,82	0,80	0,80	0,78	0,75	0,72	0,68	0,91	0,86	0,86	0,98	1,00
2	0,338	1,01	0,80	0,79	0,80	0,82	0,80	0,79	0,75	0,70	0,93	0,98	1,08	1,10
1	0,340	1,00	0,87	0,86	0,86	0,87	0,86	0,86	0,83	0,81	0,95	0,97	1,03	1,07
4-7	0,338	1,12	0,76	0,76	0,76	0,86	0,87	0,86	0,94	1,22	0,95	1,00	1,02	1,04
1-3	0,339	1,00	0,83	0,82	0,82	0,82	0,80	0,79	0,75	0,81	0,91	0,94	1,03	1,06
1-7	0,339	1,07	0,79	0,78	0,78	0,85	0,84	0,83	0,86	1,04	0,93	0,97	1,02	1,05

Column 1 shows the RMSE of the naive forecast. The remaining columns show the RMSE relative to the naive forecast. The final rows show the relative average RMSE over forecasts 1 to 4, 5 to 7, and 1 to 7, respectively. The naive forecast is based on a random walk with drift. AR(1) denotes the forecast from a first-order autoregressive model. All parameters are estimated recursively.

Figure 1: Out-of sample RMSE

(2001 Q 2 - 2007 Q4)

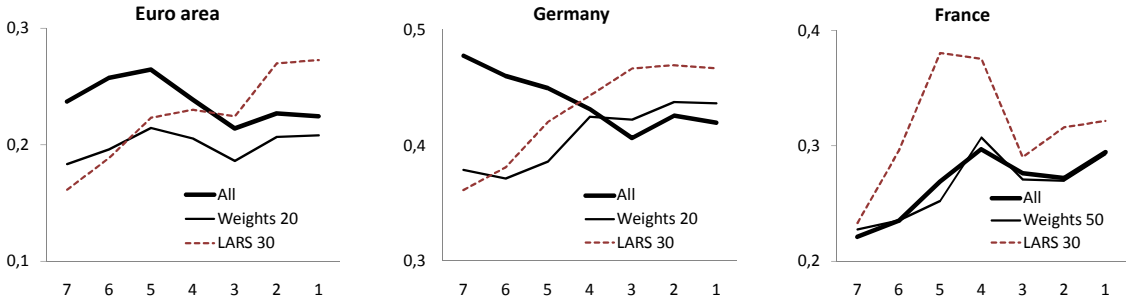

Table 3: In-sample RMSE

(1993 Q3 - 2000 Q4)

Horizon	Naive	AR(1)	Kalman filter weights								LARS			
			All	60	50	40	30	20	15	10	30	20	15	10
Euro area														
4-7	0,439	0,98	0,85	0,87	0,89	0,91	0,93	0,84	0,83	0,79	0,81	0,70	0,76	0,79
1-3	0,454	0,91	0,76	0,76	0,77	0,78	0,79	0,80	0,83	0,82	0,78	0,84	0,91	0,92
1-7	0,445	0,95	0,81	0,82	0,84	0,85	0,87	0,82	0,83	0,80	0,80	0,76	0,83	0,85
Germany														
4-7	0,663	1,29	0,95	0,97	0,97	0,95	0,93	0,90	0,91	0,92	0,92	0,90	0,89	1,03
1-3	0,666	1,18	0,91	0,91	0,92	0,94	0,96	0,94	0,96	0,97	0,96	0,95	0,95	0,96
1-7	0,665	1,24	0,93	0,95	0,95	0,95	0,94	0,92	0,93	0,94	0,94	0,92	0,92	1,00
France														
4-7	0,490	0,96	0,80	0,80	0,81	0,81	0,80	0,82	0,85	1,08	0,77	0,73	0,65	0,70
1-3	0,504	0,91	0,78	0,77	0,77	0,79	0,79	0,78	0,81	0,91	0,82	0,81	0,75	0,79
1-7	0,496	0,93	0,79	0,79	0,79	0,80	0,80	0,81	0,83	1,01	0,79	0,65	0,69	0,74

See Table 2 for explanations.

Table 4: Fixed factor model specifications

(2001 Q2 - 2007 Q4)

Horizon	Naive	AR(1)	Kalman filter weights								LARS			
			All	60	50	40	30	20	15	10	30	20	15	10
Euro area														
4-7	0,285	0,87	0,88	0,86	0,80	0,79	0,75	0,73	0,75	0,79	0,67	0,78	0,76	0,88
1-3	0,288	0,96	0,77	0,78	0,77	0,76	0,77	0,70	0,73	0,76	0,76	0,97	0,98	1,03
1-7	0,286	0,91	0,83	0,82	0,79	0,78	0,76	0,71	0,74	0,77	0,71	0,86	0,86	0,94
Germany														
4-7	0,476	1,00	0,95	0,98	0,98	0,95	0,92	0,81	0,82	0,81	0,88	0,91	0,94	1,00
1-3	0,480	1,01	0,87	0,88	0,88	0,90	0,89	0,87	0,90	0,90	0,96	0,98	1,00	0,99
1-7	0,478	1,01	0,92	0,94	0,94	0,93	0,91	0,83	0,86	0,84	0,91	0,94	0,96	1,00
France														
4-7	0,338	1,12	0,76	0,76	0,76	0,76	0,83	0,84	0,85	0,87	0,94	1,02	0,99	1,01
1-3	0,339	1,00	0,83	0,82	0,82	0,85	1,01	0,98	0,97	0,96	0,98	1,02	1,05	1,05
1-7	0,339	1,07	0,79	0,78	0,78	0,80	0,90	0,90	0,90	0,91	0,96	1,02	1,02	1,03

See Table 2 for explanations.

Table 5: Selections from forecast 2

(2001 Q1 - 2007 Q4)

Horizon	Naive	AR(1)	Kalman filter weights								LARS			
			All	60	50	40	30	20	15	10	30	20	15	10
Euro area														
3	0,288	0,95	0,74	0,75	0,74	0,73	0,78	0,70	0,71	0,79	1,25	1,13	1,00	0,99
2	0,288	0,95	0,79	0,79	0,79	0,80	0,81	0,79	0,81	0,86	1,06	1,04	1,01	1,00
1	0,290	0,97	0,77	0,77	0,78	0,79	0,79	0,78	0,81	0,88	1,02	1,00	1,00	1,00
1-3	0,288	0,96	0,77	0,77	0,77	0,77	0,79	0,76	0,78	0,84	1,11	1,06	1,00	1,00
Germany														
3	0,479	1,01	0,85	0,86	0,87	0,86	0,86	0,93	0,95	1,01	1,12	1,14	0,97	0,96
2	0,479	1,01	0,89	0,89	0,90	0,90	0,91	0,96	0,96	0,97	1,04	1,06	0,98	0,98
1	0,481	1,01	0,87	0,88	0,88	0,88	0,89	0,95	0,98	0,98	1,01	1,03	0,98	0,97
1-3	0,480	1,01	0,87	0,88	0,88	0,88	0,88	0,95	0,96	0,99	1,06	1,08	0,97	0,97
France														
3	0,338	1,01	0,82	0,82	0,79	0,75	0,69	0,74	0,68	0,89	1,01	0,99	0,99	0,99
2	0,338	1,01	0,80	0,81	0,82	0,83	0,72	0,79	0,69	0,71	1,01	1,00	1,01	0,99
1	0,340	1,00	0,87	0,87	0,88	0,83	0,83	0,87	0,78	0,86	1,01	1,00	1,00	0,99
1-3	0,339	1,00	0,83	0,83	0,83	0,81	0,75	0,80	0,72	0,82	1,01	1,00	1,00	0,99

See Table 2 for explanations.

References

Angelini E, Camba-Mendez G, Giannone D, Reichlin L, Rünstler G. 2008. Short-term forecasts of euro area GDP growth. ECB working paper 949.

Angelini E, Bańbura M, Rünstler G. 2010. Estimating and forecasting the euro area national accounts. Journal of Business Cycle Measurement and Analysis forthcoming.

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

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

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

Bańbura M, Modugno M. 2009. Maximum likelihood estimation of a large factor model on datasets with arbitrary pattern of missing data. ECB mimeo.

Bańbura M, Rünstler G. 2010. 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, forthcoming.

Banerjee A, Marcellino M, Masten I. 2005. Leading indicators for euro-area inflation and GDP growth. Oxford Bulletin of Economics and Statistics 67 : 785-813.

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

Caggiano G, Kapetianos G, Labhard V. 2009. Are more data always better for factor analysis? ECB working paper 1051.

Doz C, Giannone D, Reichlin L. 2005. A quasi maximum likelihood approach for large approximate dynamic factor models. CEPR Discussion Paper No. 5724.

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

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

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

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

Harvey AC, Koopman SJ. 2003. Computing observation weights for signal extraction and filtering. Journal of Economic Dynamics & Control 27 : 1317-1333.

Marcellino M, Schuhmacher C. 2008. Factor-MIDAS for now- and forecasting with ragged-edge data: a model comparison for German GDP. Economics Working Papers ECO2008/16, European University Institute.

Rünstler G, Barhoumi K, Benk S, Cristadoro R, den Reijer A, Jakataine A, Jelonek P, Rua A, Ruth K, van Nieuwenhuyze C. 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 JH, Watson MW. 1995. Implications of dynamic factor models for VAR analysis. Princeton University mimeo.

Stock JH, Watson MW. 2002. Macroeconomic forecasting using diffusion indexes. Journal of Business and Economics Statistics 20 : 147-162.

State space form

The transition equation of the model described in section 3.1 is given by

$$\begin{bmatrix} I_r & 0 & 0 & 0 & 0 \\ -\beta' & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & -1 & 1 & 0 \\ 0 & 0 & 0 & -\frac{1}{3} & 1 \end{bmatrix} \begin{bmatrix} f_{t+1} \\ y_{t+1} \\ y_t \\ y_{t+1}^{(3)} \\ q_{t+1} \end{bmatrix} = \begin{bmatrix} 0 \\ \mu \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} A_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \phi_t \end{bmatrix} \begin{bmatrix} f_t \\ y_t \\ y_{t-1} \\ y_t^{(3)} \\ q_t \end{bmatrix} + \begin{bmatrix} B\eta_t \\ \varepsilon_{t+1} \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

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

$$q_t = \phi_{t-1} q_{t-1} + \frac{1}{3} y_t^{(3)} \ ,$$

where $\phi_{t-1} = 0$ in the 1^{st} month and $\phi_{t-1} = 1$ 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 KF Weights	Ranking LARS
1	Index of notional stock - Money M1	2	2		13
2	Index of notional stock - Money M2	2	2		25
3	Index of notional stock - Money M3	2	2		
4	Index of Loans	2	2		11
5	ECB Nominal effective exch. rate	0	2	5	
6	ECB Real effective exch. rate CPI deflated	0	2	4	
7	ECB Real effective exch. rate producer prices deflated	0	2	30	
8	Exch. rate: USD/EUR	0	2	44	
9	Exch. rate: GBP/EUR	0	2	34	
10	Exch. rate: YEN/EUR	0	2	49	23
11	World market prices of raw materials in Euro, total, HWWA	0	2	38	
12	World market prices of raw materials in Euro, total, excl energy, HWWA	0	2		
13	World market prices, crude oil, USD, HWWA	0	2	43	
14	Gold price, USD, fine ounce	0	2	57	19
15	Brent Crude, 1 month fwd, USD/BBL converted in euro	0	2	56	
16	Retail trade, except of motor vehicles and motorcycles	2	2		6
17	IP-Total industry	3	2	37	1
18	IP-Total Industry (excl construction)	2	2	41	
19	IP-Manufacturing	2	2	20	
20	IP-Construction	3	2		
21	IP-Total Industry excl construction and MIG Energy	2	2	21	
22	IP-Energy	2	2		30
23	IP-MIG Capital Goods Industry	2	2	46	
24	IP-MIG Durable Consumer Goods Industry	2	2	22	12
25	IP-MIG Energy	2	2		
26	IP-MIG Intermediate Goods Industry	2	2	19	29
27	IP-MIG Non-durable Consumer Goods Industry	2	2	47	
28	IP-Manufacture of basic metals	2	2	58	17
29	IP-Manufacture of chemicals and chemical products	2	2	48	
30	IP-Manufacture of electrical machinery and apparatus	2	2	23	4
31	IP-Manufacture of machinery and equipment	2	2	51	9
32	IP-Manufacture of pulp, paper and paper products	2	2	52	
33	IP-Manufacture of rubber and plastic products	2	2	29	
34	Industry Survey: Industrial Confidence Indicator	0	1	3	
35	Industry Survey: Production trend observed in recent months	0	1	18	
36	Industry Survey: Assessment of order-book levels	0	1	6	
37	Industry Survey: Assessment of export order-book levels	0	1	7	27
38	Industry Survey: Assessment of stocks of finished products	0	1	28	
39	Industry Survey: Production expectations for the months ahead	0	1	60	
40	Industry Survey: Employment expectations for the months ahead	0	1	16	
41	Industry Survey: Selling price expectations for the months ahead	0	1	17	28
42	Consumer Survey: Consumer Confidence Indicator	0	1	13	
43	Consumer Survey: General economic situation over last 12 months	0	1	12	
44	Consumer Survey: General economic situation over next 12 months	0	1	14	
45	Consumer Survey: Price trends over last 12 months	0	1	31	
46	Consumer Survey: Price trends over next 12 months	0	1	32	26
47	Consumer Survey: Unemployment expectations over next 12 months	0	1	15	21
48	Construction Survey: Construction Confidence Indicator	0	1	8	
49	Construction Survey: Trend of activity compared with preceding months	0	1	11	18
50	Construction Survey: Assessment of order books	0	1	9	
51	Construction Survey: Employment expectations for the months ahead	0	1	10	15
52	Construction Survey: Selling price expectations for the months ahead	0	1	26	2
53	Retail Trade Survey: Retail Confidence Indicator	0	1	45	
54	Retail Trade Survey: Present business situation	0	1		5
55	Retail Trade Survey: Assessment of stocks	0	1		
56	Retail Trade Survey: Expected business situation	0	1	42	22
57	Retail Trade Survey: Employment expectations	0	1	33	16
58	New passenger car registrations	1	2	53	3
59	Index of Employment, Manufacturing	3	2	59	
60	Index of Employment, Total Industry (excluding construction)	3	2	54	
61	Eurostoxx 500	0	2	2	24
62	Eurostoxx 325	0	2	1	
63	US S&P 500 composite index	0	2	35	10
64	US, Dow Jones, industrial average	0	2	50	
65	US, Treasury Bill rate, 3-month	0	1		
66	US Treasury notes & bonds yield, 10 years	0	1	24	
67	Money M2 in the U.S.	2	2	55	
68	US, Unemployment rate	1	1	36	
69	US, IP total excl construction	1	2		7
70	US, Employment, civilian	1	2	40	20
71	US, Production expectations in manufacturing	0	1	25	
72	US, Consumer expectations index	0	1	39	8
73	10-year government bond yield	0	1	27	14

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

Weights: 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 KF Weights	Ranking LARS
1	Index of notional stock - Money M1	2	2		22
2	Index of notional stock - Money M2	2	2		
3	Index of notional stock - Money M3	2	2		20
4	Index of Loans	2	2		
5	ECB Nominal effective exch. rate	0	2	27	
6	ECB Real effective exch. rate CPI deflated	0	2	28	
7	ECB Real effective exch. rate producer prices deflated	0	2	26	
8	Exch. rate: USD/EUR	0	2	25	4
9	Exch. rate: GBP/EUR	0	2	33	
10	Exch. rate: YEN/EUR	0	2	49	10
11	World market prices of raw materials in Euro, total, HWWA	0	2	38	
12	World market prices of raw materials in Euro, total, excl energy, HWWA	0	2	45	
13	World market prices, crude oil, USD, HWWA	0	2	39	
14	Gold price, USD, fine ounce	0	2	34	8
15	Brent Crude, 1 month fwd, USD/BBL converted in euro	0	2	37	24
16	IP-Total industry	3	2	22	1
17	IP-Total Industry (excl construction)	2	2	30	
18	IP-Manufacturing	2	2	23	
19	IP-Construction	3	2		3
20	IP-Total Industry excl construction and MIG Energy	2	2	29	
21	IP-Energy	2	2	57	29
22	IP-MIG Capital Goods Industry	2	2	35	
23	IP-MIG Durable Consumer Goods Industry	2	2	50	15
24	IP-MIG Energy	2	2		
25	IP-MIG Intermediate Goods Industry	2	2	44	
26	IP-MIG Non-durable Consumer Goods Industry	2	2	60	13
27	IP-Manufacture of basic metals	2	2	54	
28	IP-Manufacture of chemicals and chemical products	2	2		9
29	IP-Manufacture of electrical machinery and apparatus	2	2	51	30
30	IP-Manufacture of machinery and equipment	2	2	47	
31	IP-Manufacture of pulp, paper and paper products	2	2		
32	IP-Manufacture of rubber and plastic products	2	2	56	
33	Industry Survey: Industrial Confidence Indicator	0	1	5	
34	Industry Survey: Production trend observed in recent months	0	1	21	
35	Industry Survey: Assessment of order-book levels	0	1	6	
36	Industry Survey: Assessment of export order-book levels	0	1	9	
37	Industry Survey: Assessment of stocks of finished products	0	1	10	
38	Industry Survey: Production expectations for the months ahead	0	1	15	21
39	Industry Survey: Employment expectations for the months ahead	0	1	41	
40	Industry Survey: Selling price expectations for the months ahead	0	1	18	18
41	Consumer Survey: Consumer Confidence Indicator	0	1	4	
42	Consumer Survey: General economic situation over last 12 months	0	1	3	
43	Consumer Survey: General economic situation over next 12 months	0	1	2	
44	Consumer Survey: Price trends over last 12 months	0	1	13	
45	Consumer Survey: Price trends over next 12 months	0	1	12	14
46	Consumer Survey: Unemployment expectations over next 12 months	0	1	7	16
47	Construction Survey: Construction Confidence Indicator	0	1	8	
48	Construction Survey: Trend of activity compared with preceding months	0	1	17	5
49	Construction Survey: Assessment of order books	0	1	1	11
50	Construction Survey: Employment expectations for the months ahead	0	1	11	23
51	Construction Survey: Selling price expectations for the months ahead	0	1	14	
52	Retail Trade Survey: Retail Confidence Indicator	0	1	59	
53	Retail Trade Survey: Present business situation	0	1	42	
54	Retail Trade Survey: Assessment of stocks	0	1	43	
55	Retail Trade Survey: Expected business situation	0	1	31	
56	Retail Trade Survey: Employment expectations	0	1	32	2
57	New passenger car registrations	1	2		25
58	Index of Employment, Construction	3	2		7
59	Index of Employment, Manufacturing	3	2		19
60	Index of Employment, Total Industry	3	2	52	
61	Index of Employment, Total Industry (excluding construction)	3	2		12
62	Eurostoxx 500	0	2	19	
63	Eurostoxx 325	0	2	20	
64	US S&P 500 composite index	0	2	40	
65	US, Dow Jones, industrial average	0	2	36	27
66	US, Treasury Bill rate, 3-month	0	1	53	
67	US Treasury notes & bonds yield, 10 years	0	1	16	
68	Money M2 in the U.S.	2	2	46	28
69	US, Unemployment rate	1	1	58	26
70	US, IP total excl construction	1	2		
71	US, Employment, civilian	1	2		17
72	US, Production expectations in manufacturing	0	1	48	6
73	US, Consumer expectations index	0	1	55	
74	10-year government bond yield	0	1	24	

Transformation code: 1 = 3-month difference, 2 = 3-month growth rate

Weights: 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 KF Weights	Ranking LARS
1	Index of notional stock - Money M1	2	2		17
2	Index of notional stock - Money M2	2	2	56	
3	Index of notional stock - Money M3	2	2	51	
4	Index of Loans	2	2		6
5	ECB Nominal effective exch. rate	0	2	20	
6	ECB Real effective exch. rate CPI deflated	0	2	21	
7	ECB Real effective exch. rate producer prices deflated	0	2	44	
8	Exch. rate: USD/EUR	0	2	39	16
9	Exch. rate: GBP/EUR	0	2	35	
10	Exch. rate: YEN/EUR	0	2	47	28
11	World market prices of raw materials in Euro, total, HWWA	0	2	29	
12	World market prices of raw materials in Euro, total, excl energy, HWWA	0	2	19	
13	World market prices, crude oil, USD, HWWA	0	2		
14	Gold price, USD, fine ounce	0	2		30
15	Brent Crude, 1 month fwd, USD/BBL converted in euro	0	2	28	
16	IP-Total industry	3	2	32	
17	IP-Total Industry (excl construction)	2	2	23	1
18	IP-Manufacturing	2	2	22	
19	IP-Construction	3	2		12
20	IP-Total Industry excl construction and MIG Energy	2	2	27	
21	IP-Energy	2	2		
22	IP-MIG Capital Goods Industry	2	2		
23	IP-MIG Durable Consumer Goods Industry	2	2		29
24	IP-MIG Energy	2	2	59	11
25	IP-MIG Intermediate Goods Industry	2	2	24	2
26	IP-MIG Non-durable Consumer Goods Industry	2	2	50	27
27	IP-Manufacture of basic metals	2	2	46	
28	IP-Manufacture of chemicals and chemical products	2	2	54	20
29	IP-Manufacture of electrical machinery and apparatus	2	2	60	4
30	IP-Manufacture of machinery and equipment	2	2	37	9
31	IP-Manufacture of pulp, paper and paper products	2	2	38	
32	IP-Manufacture of rubber and plastic products	2	2	48	22
33	Industry Survey: Industrial Confidence Indicator	0	1	5	
34	Industry Survey: Production trend observed in recent months	0	1	10	
35	Industry Survey: Assessment of order-book levels	0	1	4	
36	Industry Survey: Assessment of export order-book levels	0	1	9	26
37	Industry Survey: Assessment of stocks of finished products	0	1	17	15
38	Industry Survey: Production expectations for the months ahead	0	1	7	
39	Industry Survey: Employment expectations for the months ahead	0	1	36	
40	Industry Survey: Selling price expectations for the months ahead	0	1	34	
41	Consumer Survey: Consumer Confidence Indicator	0	1	6	
42	Consumer Survey: General economic situation over last 12 months	0	1	8	
43	Consumer Survey: General economic situation over next 12 months	0	1	45	
44	Consumer Survey: Price trends over last 12 months	0	1	58	
45	Consumer Survey: Price trends over next 12 months	0	1	42	
46	Consumer Survey: Unemployment expectations over next 12 months	0	1	3	
47	Construction Survey: Construction Confidence Indicator	0	1	11	
48	Construction Survey: Trend of activity compared with preceding months	0	1	16	
49	Construction Survey: Assessment of order books	0	1	18	
50	Construction Survey: Employment expectations for the months ahead	0	1	31	
51	Construction Survey: Selling price expectations for the months ahead	0	1	26	5
52	Retail Trade Survey: Retail Confidence Indicator	0	1	14	
53	Retail Trade Survey: Present business situation	0	1	15	
54	Retail Trade Survey: Assessment of stocks	0	1		18
55	Retail Trade Survey: Expected business situation	0	1	12	
56	Retail Trade Survey: Employment expectations	0	1	53	25
57	New passenger car registrations	1	2	57	8
58	Unemployment rate, total	2	1	43	3
59	US, Dow Jones, industrial average	0	2	40	
60	US, Treasury Bill rate, 3-month	0	1	41	
61	US Treasury notes & bonds yield, 10 years	0	1	25	23
62	Money M2 in the U.S.	2	2		7
63	US, Unemployment rate	1	1	49	21
64	US, IP total excl construction	1	2	52	
65	US, Employment, civilian	1	2	55	13
66	US, Production expectations in manufacturing	0	1		24
67	US, Consumer expectations index	0	1	13	10
68	Eurostoxx 500	0	2	2	19
69	Eurostoxx 325	0	2	1	
70	US S&P 500 composite index	0	2	30	
71	10-year government bond yield	0	1	33	14

Transformation code: 1 = 3-month difference, 2 = 3-month growth rate

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