Maximum likelihood estimation of large factor model on datasets with missing data: forecasting euro area GDP with mixed frequency and short-history indicators.

Marta Bańbura[1] and Michele Modugno[2]

Abstract

This paper deals with the estimation of a large dynamic factor model on datasets with a general pattern of missing data. The framework allows to handle efficiently and in a fairly automatic manner sets of indicators characterized by different publication delays, frequencies and sample lengths. This can be relevant e.g. for young economies for which many indicators are compiled only since recently. We also show how the unexpected part of new data releases is related to the forecast revision. This can be used e.g. to trace the sources of the latter back to the individual series in the case of simultaneous releases. The methodology is applied to the short-term forecasting and backdating of the euro area GDP based on a large panel of monthly and quarterly data.

Keywords: Factor Models, Forecasting, Large Cross-Sections, Missing data, EM algorithm.

JEL classification: C53, E37.

The authors would like to thank Domenico Giannone, Michele Lenza and Lucrezia Reichlin for useful discussions.

The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the European Central Bank.

[1]European Central Bank and ECARES, Université Libre de Bruxelles; email: marta.banbura@ecb.int; contact author;

[2]European Central Bank and ECARES, Université Libre de Bruxelles; email: michele.modugno@ecb.int.

1 Introduction

This paper deals with the maximum likelihood estimation of a dynamic factor model for large cross-sections with a general pattern of missing data. The framework can be used for extracting information from large datasets including e.g. mixed frequency or short history indicators or for real-time applications in which data arrive with delays and in a nonsynchronous manner. We also show how the unexpected part of a data release is related to the forecast revision and illustrate how this link can be used to understand the sources of the latter in the case of simultaneous releases.

Using a large number of indicators in order to asses the current and the future state of the economy is becoming a standard practice. In this context, factor models have emerged as an effective tool for extracting information from many indicators in a parsimonious way.[1] The fundamental assumption underlying those models is that most of the co-movement of the variables in a given dataset can be summarized by only few factors. This assumption seems to be justified in the case of macroeconomic and financial data. Forecasting macroeconomic variables is one of the domains where factor models show a great potential.[2]

This paper builds on the approximate dynamic factor model of Giannone, Reichlin, and Small (2008). It has been implemented in several countries and proved to perform well in short-term forecasting, see e.g. Bańbura and Rünstler (2007) for the euro area GDP, Aastveit and Trovik (2007) for Norwegian GDP, Matheson (2007) for New Zealand GDP and inflation. It allows for explicit modelling of the dynamics of the factors[3] and for missing data at the end of the sample. This feature can be exploited e.g. to extract information from timely indicators in real-time forecasting.[4] The framework also allows for a certain degree of cross and serial correlation in the idiosyncratic components.

In Giannone, Reichlin, and Small (2008) the estimation is performed in two steps on a balanced panel.[5] To establish the relation between the GDP and the factors, the latter are extracted through the principal components analysis on monthly indicators and converted to the quarterly frequency. Therefore the methodology cannot be easily applied to general mixed frequency panels including series of different lengths. The maximum likelihood approach that we propose is able to handle large datasets with an arbitrary pattern of data availability efficiently and in a fairly automatic manner. This is particularly relevant for the euro area or other young economies for which many series have been compiled only since recently (e.g. euro

[1] See Forni, Hallin, Lippi, and Reichlin (2000) and Stock and Watson (2002a) for the theoretical foundations.

[2] See e.g. Bernanke and Boivin (2003); Boivin and Ng (2005); D'Agostino and Giannone (2006); Forni, Hallin, Lippi, and Reichlin (2005, 2003); Giannone, Reichlin, and Sala (2004); Marcellino, Stock, and Watson (2003); Stock and Watson (2002a,b).

[3] shown to be important in the simulation study of Doz, Giannone, and Reichlin (2005);

[4] in contrast to models based on balanced datasets as in e.g. Forni, Hallin, Lippi, and Reichlin (2000); Stock and Watson (2002a);

[5] The consistency has been proved by Doz, Giannone, and Reichlin (2005).

area Purchasing Managers' (PM) Surveys). Moreover as the series measured at a lower frequency can be interpreted as "high frequency" indicators with missing data, mixed frequency datasets can be easily handled. This can be important for two reasons: first, the information in the indicators sampled at a lower frequency (e.g. consumption, employment) can be used to extract the factors; second, the forecasts of the former can be easily obtained.

Apart from data availability considerations the maximum likelihood estimation is more appealing than the two-step method because it is more efficient and, more importantly, it provides framework for imposing restrictions on the parameters. For example Reis and Watson (2007) impose restrictions on the parameters of the dynamic factor model in order to estimate the pure inflation. Maximum likelihood approach has been used in the classical factor analysis for small datasets (see e.g. Geweke, 1977; Sargent and Sims, 1977; Watson and Engle, 1983). Doz, Giannone, and Reichlin (2006) show that maximum likelihood is consistent, robust and computationally feasible also in the case of large cross-sections. To maximise the likelihood over the high-dimensional parameter space they propose to use the Expectation-Maximisation (EM) algorithm. The EM algorithm was first applied for a dynamic factor model by Watson and Engle (1983) on a small cross-section. They cast the model in a state space form and derive the EM steps in the case without missing data. Shumway and Stoffer (1982) show how to implement the EM algorithm for a state space form with missing data, however only in the case in which the matrix linking the states and the observables is known. We extend those results to the case in which all the parameters of the state space form need to be estimated.

We apply the methodology to short-term forecasting of euro area GDP on the basis of large panel of monthly and quarterly indicators. GDP is an important measure of economic performance, however it is released only with about 2 months delay. In the meantime more timely indicators can provide information on the current state of the economy. For example, Giannone, Reichlin, and Small (2008) claim that the Federal Reserve Bank of Philadelphia Business Outlook Survey is crucial for nowcasting the U.S. GDP due to its early release (cf. Bańbura and Rünstler, 2007, for the case of euro area). We want to examine the effect of quarterly variables and short history monthly series of PM Surveys on the forecast. For example, the PM Surveys are considered to be important soft indicators of the real-activity in the euro area (e.g. Purchasing Managers' Index is analysed in the ECB Monthly Bulletin on a regular basis), however their short sample length is prohibitive for many models that cannot flexibly deal with missing data.

We start by comparing the maximum likelihood approach with the two-step procedure used by Giannone, Reichlin, and Small (2008) and Bańbura and Rünstler (2007) on the benchmark dataset containing 70 monthly "long" indicators dating back to at least to 1993, similar to the one used by e.g Bańbura and Rünstler (2007) or ECB (2008). In the following exercise we augment the benchmark dataset by quarterly and/or short monthly indicators and estimate

the model by maximum likelihood. We find that the results based on different methods and different datasets are comparable. This means that, on one hand, adding quarterly variables and PM Surveys does not lead to forecast accuracy improvements, but on the other hand, our methodology allows us to analyse datasets including the mixed frequency and short history indicators and obtain their forecast in a unified framework. In the following exercise we illustrate how the "news" in the consecutive releases of the industrial production, PM and European Commission Surveys revise the GDP forecast. Finally, we show that the framework can be used for backdating. In particular, the back estimates of GDP are fairly close to the true values.

Related literature includes Camacho and Perez-Quiros (2008) who estimate a small dynamic factor model by maximum likelihood to obtain the real-time estimates of the GDP from monthly indicators. Breitung and Schumacher (2008) forecast GDP from large number of monthly indicators using the EM approach, however they do not exploit the dynamics of the factors. Proietti (2008) estimates a factor model[6] from a balanced panel by EM for interpolation of expenditure and output components of the GDP and shows how to incorporate relevant accounting and temporary constraints.

As for the interpolation and backcasting, Angelini, Henry, and Marcellino (2006) propose methodology based on large cross-sections. In contrast to their procedure our method exploits the dynamics of the data and is based on maximum likelihood which allows for imposing restrictions and is more efficient for smaller cross-sections.

The paper is organized as follows. Section 2 presents the model, discusses the estimation and explains how the news content can be extracted. Section 3 describes the empirical application. Section 4 concludes. The technical details, data description and robustness checks are provided in the Appendix B.

2 The model

Let $y_t = (y_{1,t}, y_{2,t}, \ldots, y_{n,t})$ denote a stationary n-dimensional vector process standardised to mean 0 and unit variance. We assume that y_t follows an approximate factor model given by:

$$y_t = \Lambda f_t + \epsilon_t \qquad \epsilon_t \sim N(0, \Psi)\,, \tag{1}$$

where f_t is a $r \times 1$, $r \ll n$ vector of common factors and ξ_t is the idiosyncratic error, uncorrelated with f_t at all leads and lags. However, contrary to the case of the exact dynamic factor model, the idiosyncratic error can be weakly serially- and cross-correlated (cf. Doz, Giannone, and

[6] a la Watson and Engle (1983) and a la Stock and Watson (2002b);

Reichlin, 2006, for the technical details). Moreover we assume that the common factors f_t follow a stationary VAR process of order p:

$$f_t = A_1 f_{t-1} + A_2 f_{t-2} + \cdots + A_p f_{t-p} + u_t \qquad u_t \sim N(0, Q)\,. \tag{2}$$

2.1 Estimation

As pointed out in Doz, Giannone, and Reichlin (2006) it is not obvious how to model parametrically the weak cross-correlation of the idiosyncratic component in (1). However they show that the model can be estimated by quasi maximum likelihood, where the miss-specified model is the exact factor model

$$y_t = \Lambda f_t + \xi_t \qquad \xi_t \sim N(0, R)\,, \tag{3}$$

with R diagonal and ξ_t serially uncorrelated. They prove that the maximum likelihood estimates of the model given by (1) - (2) and of the miss-specified given by (3) - (2) are asymptotically equivalent. Therefore, in what follows we will derive the estimates under the miss-specified likelihood.

In the main text we set for simplicity $p = 1$, the case of $p > 1$ is discussed in the Appendix. Accordingly, the system given by (3) and (2) can be written in a state space form with the latent factors as states:

$$\begin{aligned} y_t &= \Lambda f_t + \xi_t \qquad & \xi_t &\sim N(0, R)\,, \\ f_t &= A_1 f_{t-1} + u_t \qquad & u_t &\sim N(0, Q)\,. \end{aligned} \tag{4}$$

A direct maximisation of the likelihood for (4) is computationally not feasible for large n. However, as argued in Doz, Giannone, and Reichlin (2006), the computational complexity can be circumvented by means of the Expectation-Maximisation (EM) algorithm. It offers a solution to problems for which incomplete or latent data yield the likelihood intractable. The essential idea is to write the likelihood as if the data was complete and to "fabricate" the missing data in the expectation step.[7]

Let us collect the parameters of the model in $\theta = (\Lambda, A_1, R, Q)$ and let $l(Y, F; \theta)$ denote the joint log-likelihood of y_t and f_t, $t = 1, \ldots, T$. Given the available data Ω_T, the EM algorithm converges towards the maximum likelihood estimates in a sequence of two alternating steps:

1. E-step - the expectation of the log-likelihood conditional on the data is calculated using the estimates from the previous iteration $\theta(r)$:

$$L(\theta, \theta(r)) = \mathrm{E}_{\theta(r)}\left[l(Y, F; \theta) | \Omega_T \right];$$

[7]The EM algorithm was proposed by Dempster, Laird, and Rubin (1977). For overview see e.g. McLachlan and Krishnan (1996).

2. M-step - the parameters are re-estimated through the maximisation of the expected log-likelihood with respect to θ:

$$\theta(r+1) = \arg\max_{\theta} L(\theta, \theta(r)) \,. \tag{5}$$

From the maximisation of (5) it follows that the estimates in the $r+1$ iteration are given by:

$$A_1(r+1) = \left(\sum_{t=1}^{T} \mathrm{E}_{\theta(r)}\left[f_t f_{t-1}' | \Omega_T\right]\right)\left(\sum_{t=1}^{T} \mathrm{E}_{\theta(r)}\left[f_{t-1} f_{t-1}' | \Omega_T\right]\right)^{-1}, \tag{6}$$

$$\Lambda(r+1) = \left(\sum_{t=1}^{T} \mathrm{E}_{\theta(r)}\left[y_t f_t' | \Omega_T\right]\right)\left(\sum_{t=1}^{T} \mathrm{E}_{\theta(r)}\left[f_t f_t' | \Omega_T\right]\right)^{-1}, \tag{7}$$

$$Q(r+1) = \frac{1}{T}\left(\sum_{t=1}^{T} \mathrm{E}_{\theta(r)}\left[f_t f_t' | \Omega_T\right] - A_1(r+1)\sum_{t=1}^{T} \mathrm{E}_{\theta(r)}\left[f_{t-1} f_t' | \Omega_T\right]\right) \tag{8}$$

and

$$R(r+1) = \mathrm{diag}\left(\frac{1}{T}\left(\sum_{t=1}^{T} \mathrm{E}_{\theta(r)}\left[y_t y_t' | \Omega_T\right] - \Lambda(r+1)\sum_{t=1}^{T} \mathrm{E}_{\theta(r)}\left[f_t y_t' | \Omega_T\right]\right)\right). \tag{9}$$

Watson and Engle (1983) derive the estimates when y_t does not contain missing data. In that case we have

$$\mathrm{E}_{\theta(r)}\left[y_t y_t' | \Omega_T\right] = y_t y_t' \quad \text{and} \quad \mathrm{E}_{\theta(r)}\left[y_t f_t' | \Omega_T\right] = y_t \mathrm{E}_{\theta(r)}\left[f_t' | \Omega_T\right] \tag{10}$$

and $\mathrm{E}_{\theta(r)}\left[f_t | \Omega_T\right]$, $\mathrm{E}_{\theta(r)}\left[f_t f_t' | \Omega_T\right]$ and $\mathrm{E}_{\theta(r)}\left[f_t f_{t-1}' | \Omega_T\right]$ can be obtained through the Kalman smoother for the state space representation (4) (with the parameters $\theta(r)$). Shumway and Stoffer (1982) provide the estimates also for the incomplete data set however in the case when Λ is known. We provide the EM steps for the general case when Λ is unknown and y_t contains some missing values. In that case (10) no longer holds. The formulas (6) and (8) remain unaffected, however (7) and (9) need to be modified as follows:

$$\mathrm{vec}\big(\Lambda(r+1)\big) = \left(\sum_{t=1}^{T} \mathrm{E}_{\theta(r)}\left[f_t f_t' | \Omega_T\right] \otimes W_t\right)^{-1} \mathrm{vec}\left(\sum_{t=1}^{T} W_t y_t \mathrm{E}_{\theta(r)}\left[f_t' | \Omega_T\right]\right)$$

and

$$\begin{aligned} R(r+1) &= \mathrm{diag}\Bigg(\frac{1}{T}\sum_{t=1}^{T}\Big(W_t y_t y_t' W_t' - W_t y_t \mathrm{E}_{\theta(r)}\left[f_t' | \Omega_T\right]\Lambda(r+1)' W_t - W_t \Lambda(r+1)\mathrm{E}_{\theta(r)}\left[f_t | \Omega_T\right] y_t' W_t' \\ &+ W_t \Lambda(r+1)\mathrm{E}_{\theta(r)}\left[f_t f_t' | \Omega_T\right]\Lambda(r+1)' W_t + (I - W_t) R(r) (I - W_t)\Big)\Bigg), \end{aligned}$$

where W_t is a diagonal matrix of size n with i^{th} diagonal element equal to 0 if $y_{i,t}$ is missing and equal to 1 otherwise. The derivations of the formulas are provided in the Appendix along with the results of simulations similar to those in Doz, Giannone, and Reichlin (2006) that verify the convergence of the estimates for different portions of missing data.

2.2 Forecasting

Given the estimates of the parameters $\hat{\theta} = (\hat{\Lambda}, \hat{A}_1, \hat{R}, \hat{Q})$ and the data set Ω_T the best linear predictions: $\mathbb{P}(y_{T+h}|\Omega_T)$, $h \geq 0$ can be obtained from the Kalman filter. In case some of the observations in y_t are missing, the corresponding rows in y_t and $\hat{\Lambda}$ (and the corresponding rows and columns in $\hat{R}$) are skipped when applying the Kalman filter (cf. Durbin and Koopman, 2001). The interpolates and backdata ($h < 0$) can be also simply obtained from the Kalman smoother.

To deal with mixed frequencies, low-frequency series are treated as high-frequency variables with missing data.

2.3 News and forecast revisions

In this section we show how the unexpected content, i.e. the *news*, in a data release is linked to the resulting forecast revision.

Let Ω_{v-1} and Ω_v be two consecutive vintages of data, consequently $\Omega_{v-1} \subset \Omega_v$.[8] Let I_v denote the *news* in Ω_v with respect to Ω_{v-1}. For example, let us assume that the difference between Ω_{v-1} and Ω_v is the release of the European Commission (EC) Surveys for the period t_i. The *news* is $I_v = y_{t_i}^{EC} - \mathbb{P}\left(y_{t_i}^{EC}|\Omega_{v-1}\right)$, where $y_{t_i}^{EC}$ is the vector of the new observations.

Assume that we are interested in how this *news* revises the GDP forecast for the period t_j. As $I_v \perp \Omega_{v-1}$ we can write

$$\mathbb{P}\left(\cdot|\Omega_v\right) = \mathbb{P}\left(\cdot|\Omega_{v-1}\right) + \mathbb{P}\left(\cdot|I_v\right)$$

or

$$\underbrace{\mathbb{P}\left(y_{t_j}^{GDP}|\Omega_v\right)}_{\text{new forecast}} = \underbrace{\mathbb{P}\left(y_{t_j}^{GDP}|\Omega_{v-1}\right)}_{\text{old forecast}} + \mathbb{P}\Big(y_{t_j}^{GDP}|\underbrace{I_v}_{\text{news}}\Big)\,.$$

In other words, the updated forecast can be decomposed into the sum of the old forecast and of the contribution from the *news* in the latest release.

To compute the latter we use the fact that

$$\mathbb{P}\left(y_{t_j}^{GDP}|I_v\right) = \mathrm{E}\left(y_{t_j}^{GDP} I_v'\right) \mathrm{E}\left(I_v I_v'\right)^{-1} I_v\,.$$

Furthermore, given the model (4) we can write

$$\begin{aligned} y_{t_j}^{GDP} &= \lambda_{GDP} f_{t_j} + \xi_{t_j}^{GDP}\,, \\ I_v &= y_{t_j}^{EC} - y_{t_j|\Omega_{v-1}}^{EC} = \lambda_{EC}\left(f_{t_j} - f_{t_j|\Omega_{v-1}}\right) + \xi_{t_j}^{EC}\,, \end{aligned}$$

[8] In what follows, we do not take into account the revisions and changes in the parameter estimates. The influence of those factors needs to be analysed separately.

where λ_{GDP} and λ_{EC} are the rows of Λ corresponding to GDP and EC Surveys, respectively. It can be shown (see the Appendix) that:

$$\begin{aligned} \mathrm{E}\left(y_{t_j}^{GDP} I_v'\right) &= \lambda_{GDP} \mathrm{E}(f_{t_j} - f_{t_j|\Omega_{v-1}})(f_{t_i} - f_{t_i|\Omega_{v-1}})'\lambda_{EC}' \quad \text{and} \\ \mathrm{E}\left(I_v I_v'\right) &= \lambda_{EC} \mathrm{E}(f_{t_i} - f_{t_i|\Omega_{v-1}})(f_{t_i} - f_{t_i|\Omega_{v-1}})'\lambda_{EC}' + R_{EC}\,, \end{aligned}$$

where R_{EC} is a diagonal matrix with elements of R corresponding to the EC Surveys. The expectations $\mathrm{E}(f_{t_j} - f_{t_j|\Omega_{v-1}})(f_{t_i} - f_{t_i|\Omega_{v-1}})'$ and $\mathrm{E}(f_{t_i} - f_{t_i|\Omega_{v-1}})(f_{t_i} - f_{t_i|\Omega_{v-1}})'$ can be obtained from the Kalman filter, see the Appendix.

Consequently, we can find a vector B such that the following holds:

$$\underbrace{y_{t_j|\Omega_v}^{GDP}}_{\text{new forecast}} = \underbrace{y_{t_j|\Omega_{v-1}}^{GDP}}_{\text{old forecast}} + B\Big(\underbrace{y_{t_i}^{EC} - y_{t_i|\Omega_{v-1}}^{EC}}_{\text{news}}\Big)\,. \tag{11}$$

This enables us trace the sources of forecast revisions.[9] More precisely, in the case of a simultaneous release of several (groups of) variables it is possible to decompose the resulting forecast revision into contributions from the *news* in individual (groups of) series, see the illustration in Section 3.3.[10] In addition, we can produce statements like e.g. "after the release of the EC Surveys, the forecast of GDP went up because the indicators turned out to be (on average) higher than expected".[11]

3 Forecasting the euro area GDP

In the empirical part of the paper we apply the methodology described above in the context of short-term forecasting and backcasting of the euro area GDP. In particular, we analyse the role of quarterly variables and the Purchasing Managers' Surveys.

3.1 Data

The dataset contains in total 114 indicators and was downloaded in September 2007 directly after the release of industrial production series. The detailed description including list of the series, their availability and applied transformations is provided in the Appendix.

For the purpose of the empirical exercise the dataset is divided into three categories. The first group consists of 70 indicators similar to the one described in Bańbura and Rünstler (2007)

[9] Note, that the contribution from the *news* is equivalent to the change in the overall contribution of the series to the forecast (the measure proposed in Bańbura and Rünstler, 2007) when the correlations between the predictors are not exploited in the model. Otherwise, those measures are different. In particular, there can be a change in the overall contribution of a variable even if no new information on this variable was released.

[10] If the release concerns only one group or one series, the contribution of its *news* is simply equal to the change in the forecast.

[11] This holds of course for the indicators with positive entries in B.

or used at the ECB for the short term forecast, cf. ECB (2008). It contains information on world prices, trade, industrial production, European Commission Surveys (EC Surveys), retail trade, labour, financial, US and interest rates. All the series in this group date back at least to January 1993, but most of them have a longer history. We refer to this group as the *Benchmark* dataset.

The second group contains 16 monthly "short-history" indicators, namely the Purchasing Managers' Surveys (PM Surveys). They are available from August 1997 or July 1998. We refer to this group of data as *Short Monthly* dataset.

Finally the *Quarterly* dataset includes quarterly variables (with different time spans) on employment, unit labour cost, hourly labour cost, real GDP, real value added and EC Survey on capacity utilisation.

Following Giannone, Reichlin, and Small (2008) the monthly indicators are transformed so as to be in line with the nature of the quarterly variables (which are differences of 3 month totals).

3.2 Forecast evaluation

Given the trade-off between the timeliness and the information content, in the evaluation we want to account for the differences in the publication delays. In this, we aim at replicating as closely as possible the real-time forecasting exercise. As the real-time vintages are not available for all the variables of interest and whole evaluation period, we perform the so called pseudo-real time exercise. It means that we use the final figures as of mid-September 2007 however at each point of the evaluation sample we apply appropriate publication lags following the availability pattern for mid-September (this relies on the assumption that the release dates do not change much from month to month). For example, in mid-September the last available figure on the industrial production is for July. Consequently when we evaluate the model in e.g. April the data for the industrial production "ends" in February. The same mechanism is applied to all the variables. The procedure for quarterly variables follows a similar logic modified to take into account the quarterly frequency of the releases.

To deal with mixed frequency data it is assumed that y_t are observed at a monthly frequency and the quarterly indicators are to be understood as monthly series with missing values in the first and second month of each quarter. Consequently, the GDP figure for a given quarter is assigned to the last month of this quarter.

We evaluate forecast accuracy for different forecast horizons: 0- (nowcast), 1- and 2-quarters ahead. In the case of nowcast we project the current quarter. This is relevant as the euro area GDP is released only around 6 weeks after the end of the respective quarter. Moreover, the available information changes depending on the respective month within a given quarter.

Therefore we evaluate separately forecasts made in the first, second or third month of a quarter. For example, 1-quarter-ahead forecast made in the second month means that we make forecast for the second quarter relying on the information available in February[12], for the third quarter using the information available in May, etc.

All the results in the paper correspond to a model with two factors ($r = 2$).[13]

The evaluation period is eight years, going from 1999Q3 to 2007Q2. For the measure of prediction accuracy we choose the mean squared forecast error (MSFE). All the tables present MSFEs relative to the a naïve constant growth model.[14]

The estimation sample starts in July 1993. We choose a recursive estimation which means that the sample length increases each time that more information becomes available.

We start by comparing the maximum likelihood method (ML) with the two-step approach (2S) proposed by Doz, Giannone, and Reichlin (2005) and used currently at the ECB for the short-term forecasting. One of the differences between the two methodologies applied on the benchmark dataset is in the way in which GDP is included in the analysis. In the 2S approach, the GDP is projected on the factors bridged to quarterly frequency and, given the parameters and the factor forecast produced by the factor model, the forecast of the GDP growth is computed. In the ML case the GDP is added to the benchmark dataset. This means that the GDP series will be the only one with missing values not only at the end of sample, but also in the first two months of each quarter.

Table 1: MSFEs for maximum likelihood and two-step approach

	Nowcast		1-quarter-ahead		2-quarter-ahead	
month	2S	ML	2S	ML	2S	ML
1	0.63	0.64	0.83	0.83	0.94	0.96
2	0.60	0.61	0.74	0.75	0.89	0.93
3	0.59	0.58	0.68	0.68	0.89	0.90

Notes: comparison between the MSFEs ratios to the constant growth rate model of the nowcast, 1-quarter-ahead and 2-quarter-ahead forecasts produced with a factor model estimated on the benchmark dataset with the maximum likelihood (ML) and the two-step (2S) approach.

Table 1 reports the corresponding MSFEs for the two methods. We can see that both produce comparable results. ML, however, has the advantage that it is easier to implement as it

[12] i.e. second month of first quarter;

[13] The results are qualitatively the same for smaller/larger number of factors.

[14] Let X_t and x_t denote the log of GDP and its quarterly growth rate, respectively (observed at quarterly frequency). The naïve constant growth model implies that X_t is a random walk with drift: $X_{t+1} = \delta + X_t + \epsilon_{t+1}$. Consequently $x_{t+1} = \delta + \epsilon_{t+1}$ and the optimal predictor is the average of the past growth rates and it does not depend on the forecast horizon h:

$$\hat{x}_{t+h|t} = \hat{\delta} = \frac{1}{t - T_0 + 1} \sum_{k=T_0}^{t} \hat{x}_k \, .$$

produces automatically the forecast for the GDP without the need for the second step.

In the following exercise we want to explore the effects of adding data with shorter history or quarterly frequency on the forecast accuracy. Table 2 presents the results for the nowcast and 1- and 2-quarter-ahead forecasts, respectively. In all the cases the parameters are estimated by maximum likelihood, what differs is the composition of the datasets. The first column reports the results obtained on the benchmark dataset (Bench); the second column reports the results obtained on the benchmark dataset augmented by the short monthly PM Surveys data (B+Monthly); the third column corresponds to the benchmark dataset plus all the quarterly variables (B+Quarterly); finally the fourth column reports the results obtained using all the data together (All). Figure 1 shows the true GDP growth and the nowcasts obtained with the four different datasets. The naïve "constant growth" projection is also plotted.

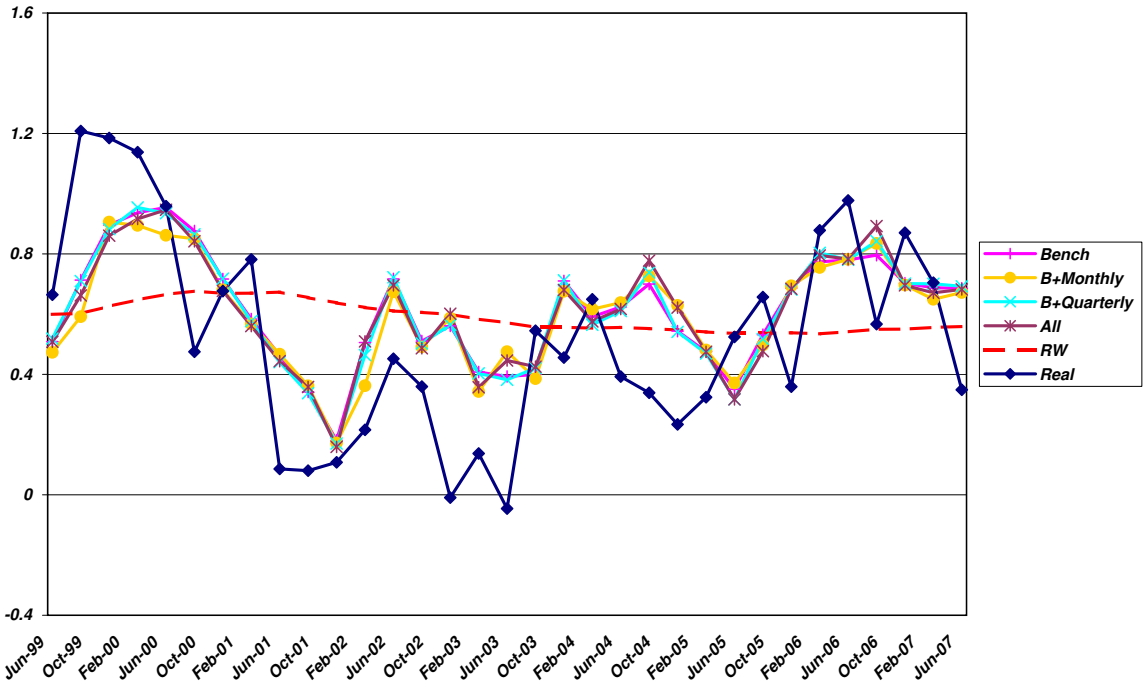
Table 2: MSFEs for maximum likelihood approach with different datasets

month	Bench	B+Monthly	B+Quarterly	All
Nowcast				
1	0.64	0.67	0.63	0.67
2	0.61	0.63	0.60	0.62
3	0.58	0.61	0.58	0.65
1-quarter-ahead forecast				
1	0.83	0.85	0.83	0.85
2	0.75	0.77	0.75	0.77
3	0.68	0.70	0.67	0.71
2-quarter-ahead forecast				
1	0.96	1.07	0.96	1.06
2	0.93	0.98	0.92	0.97
3	0.90	0.90	0.90	0.89

Notes: MSFEs ratios with respect to the constant growth rate model of the nowcast and forecasts produced with a factor model estimated on the benchmark dataset (Bench), on the benchmark dataset plus the short monthly indicators, on the benchmark dataset plus the quarterly indicators (B+Quarterly) and on the benchmark dataset plus the short monthly and the quarterly indicators (All).

We can see that the accuracy of the projections based on different datasets is comparable. Moreover, as Figure 1 shows for the nowcast, the differences between the point estimates are very small. In other words, there are no gains in forecast accuracy from augmenting the dataset by the additional indicators. On the other hand, the accuracy of forecasts based on the extended datasets does not deteriorate (perhaps with the exception of 2-quarter-ahead forecast). Consequently, we can include the additional series in the dataset and obtain their forecasts (or back estimates). This also allows us to analyse the *news* they provide, see the next section.

Figure 1: Nowcast of the GDP



3.3 News and forecast revisions, illustration

Let us recall from Section 2.3 that the *news* is understood as the unexpected part of a data release.

In the following exercise, on the example of the fourth quarter of 2001, we illustrate how the projection evolves with new data releases and what are the contributions of the *news* in different groups of series to the forecast revisions. We use the benchmark dataset augmented by the PM Surveys and we divide the indicators into 4 groups: industrial production indicators (IP), EC and PM Surveys and Other. We start to forecast in June 2001 (corresponding to the 2-quarter-ahead forecast in the third month) and we revise the projection each month as new data arrive. The last estimate is obtained in February 2002 and the actual GDP for the fourth quarter is released in March. Note that the 2 last projections are actually "backcasts" - they refer to the previous quarter. At each step we break down the forecast revision into the contributions of the *news* from different data groups using the formula (11).[15]

Figure 2 shows the evolution of the forecast as the new information arrives, the actual value of the GDP for the fourth quarter and the decomposition of the revisions.[16] On average, the biggest contributions to the revisions come from the EC Surveys. In contrast, those from the industrial production have a sizeable effect on the projection only in the case of the backcast. This confirms the results of Bańbura and Rünstler (2007) on the important role of soft data for the GDP projections when the hard data for the relevant periods are not yet available.

Another observation is that the EC and PM Surveys do not always carry the same information. There are cases when their contributions are of the same sign and the cases when the opposite holds. In particular, in the case of nowcast the *news* in the EC Surveys "moves" the projection "towards" the actual value. This is not always the case for the PM Surveys.

3.4 Backdating the GDP

A useful feature of our framework is that the estimates of the missing observations in the panel can be obtained from the Kalman smoother in an automatic manner. This enables us to use it e.g. in order to backdate a short history series or to interpolate quarterly variables. In this section we evaluate the model at backdating the GDP. We include the GDP quarterly growth rate only as of June 2000 and discard all the previous observations. We run the model on the four datasets used in the forecast evaluation. Figure 3 plots the back estimates of the GDP from different datasets, the estimate obtained with a constant growth rate model and the actual quarterly growth rate of the GDP. Table 3 reports the mean squared errors of the

[15] The contribution of a group of series is the sum of the contributions of the series within this group.
[16] In fact "Other" contains also the effect of re-estimation.

Figure 2: Contribution of news to forecast revisions for 2001Q4

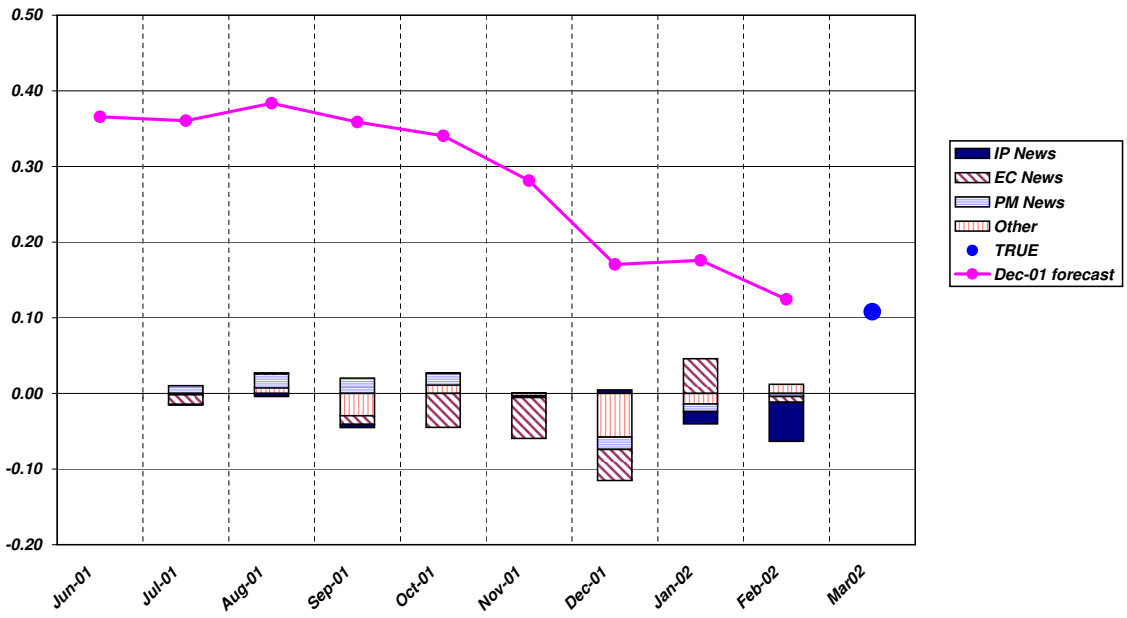
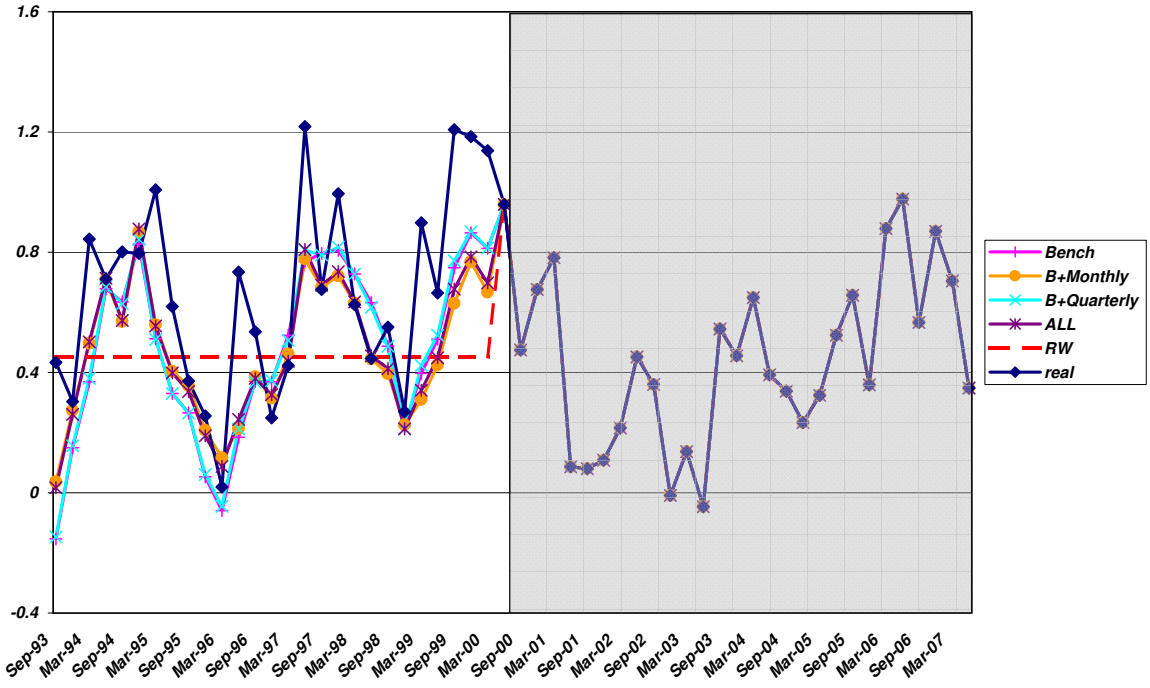


Figure 3: Back estimates of the GDP



backdated GDP obtained with the four different datasets relative to the naïve constant growth rate estimate.

Table 3: MSEs of back estimates of the GDP

Bench	B+Monthly	B+Quarterly	ALL
0.59	0.58	0.55	0.53

Notes: MSEs relative to the constant growth rate model of the backcast produced with a factor model estimated on the benchmark dataset (Bench) including the "trimmed" GDP, on the benchmark dataset augmented by the short monthly indicators (B+Monthly), on the benchmark dataset plus the quarterly indicators (B+Quarterly) and on the entire dataset (All).

As we can see from Table 3 and Figure 3, independently of the dataset used, the backdating seems to capture the movements of the GDP, giving reasonable estimates of the past values of the series. Again, different datasets yield comparable results.

4 Summary

This paper proposes a methodology for the estimation of a large dynamic factor model for datasets with arbitrary pattern of missing values. For that we adopt the maximum likelihood approach, which is shown by Doz, Giannone, and Reichlin (2005) to be consistent as well as computationally feasible for large cross-sections once the Expectation-Maximisation (EM) algorithm is applied. We show how the steps of the EM algorithm should be modified in the case of missing data. In addition, we derive the link between the unexpected component of data releases and the resulting forecast revision.

We apply this methodology for the short-term forecasting of the euro area GDP on the basis of large monthly dataset. Thanks to the flexibility of the framework in dealing with missing data, short history and quarterly variables can be also considered (e.g. Purchasing Managers' Surveys, GDP components or labour statistics). The effect of including these indicators in the large monthly dataset similar to the one used in e.g. Bańbura and Rünstler (2007) or ECB (2008) is evaluated in an out-of-sample forecast exercise. The results indicate that the additional indicators do not improve the precision of the projections. On the other hand, they can be analysed and forecast in a single model. Finally, we show that the framework can be easily used for back estimation. In particular, the back estimates of the GDP are fairly close to the true values.

References

Aastveit, K., and T. Trovik (2007): "Nowcasting Norwegian GDP: The role of asset prices in a small open economy," Norges Bank Working Paper 2007/09.

Angelini, E., J. Henry, and M. Marcellino (2006): "Interpolation and backdating with a large information set," *Journal of Economic Dynamics and Control*, 30(12), 2693–2724.

Bańbura, M., and G. Rünstler (2007): "A look into the factor model black box. Publication lags and the role of hard and soft data in forecasting GDP.," Working Paper Series 751, European Central Bank.

Bernanke, B. S., and J. Boivin (2003): "Monetary policy in a data-rich environment," *Journal of Monetary Economics*, 50(3), 525–546.

Boivin, J., and S. Ng (2005): "Understanding and Comparing Factor-Based Forecasts," *International Journal of Central Banking*, 3, 117–151.

Breitung, J., and C. Schumacher (2008): "Real-time forecasting of German GDP based on a large factor model with monthly and quarterly Data," *International Journal of Forecasting*, 24, 386–398.

Camacho, M., and G. Perez-Quiros (2008): "Introducing the EURO-STING: Short Term INdicator of Euro Area Growth," Banco de España Working Papers 0807, Banco de España.

D'Agostino, A., and D. Giannone (2006): "Comparing alternative predictors based on large-panel factor models," Working Paper Series 680, European Central Bank.

Dempster, A., N. Laird, and D. Rubin (1977): "Maximum Likelihood Estimation From Incomplete Data," *Journal of the Royal Statistical Society*, 14, 1–38.

Doz, C., D. Giannone, and L. Reichlin (2005): "A two-step estimator for large approximate dynamic factor models based on Kalman filtering," Manuscript, Université Libre de Bruxelles.

——— (2006): "A Quasi Maximum Likelihood Approach for Large Approximate Dynamic Factor Models," Working Paper Series 674, European Central Bank.

Durbin, J., and S. J. Koopman (2001): *Time Series Analysis by State Space Methods*. Oxford University Press.

ECB (2008): "Short-term forecasts of economic activity in the euro area," in *Monthly Bulletin, April*, pp. 69–74. European Central Bank.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000): "The Generalized Dynamic Factor Model: identification and estimation," *Review of Economics and Statistics*, 82, 540–554.

——— (2003): "Do Financial Variables Help Forecasting Inflation and Real Activity in the Euro Area?," *Journal of Monetary Economics*, 50, 1243–55.

——— (2005): "The Generalized Dynamic Factor Model: one-sided estimation and forecasting," *Journal of the American Statistical Association*, 100, 830–840.

Geweke, J. F. (1977): "The Dynamic Factor Analysis of Economic Time Series Models," in *Latent Variables in Socioeconomic Models*, ed. by D. Aigner, and A. Goldberger, pp. 365–383. North-Holland.

Giannone, D., L. Reichlin, and L. Sala (2004): "Monetary Policy in Real Time," in *NBER Macroeconomics Annual*, ed. by M. Gertler, and K. Rogoff, pp. 161–200. MIT Press.

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.

Jungbacker, B., and S. J. Koopman (2008): "Likelihood-based Analysis for Dynamic Factor Models," Tinbergen Institute Discussion Papers 08-007/4, Tinbergen Institute.

Koopman, S. J., and J. Durbin (2000): "Fast filtering and smoothing for multivariate state space models," *Journal of Time Series Analysis*, 21(3), 281–296.

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

Matheson, T. (2007): "An analysis of the informational content of New Zealand data releases: the importance of business opinion surveys," Reserve Bank of New Zealand Discussion Paper 2007/13.

McLachlan, G. J., and T. Krishnan (1996): *The EM Algorithm and Extensions*. John Wiley and Sons.

Proietti, T. (2008): "Estimation of Common Factors under Cross-Sectional and Temporal Aggregation Constraints: Nowcasting Monthly GDP and its Main Components," Manuscript.

Reis, R., and M. W. Watson (2007): "Relative Goods' Prices and Pure Inflation," NBER Working Paper 13615.

Sargent, T. J., and C. Sims (1977): "Business Cycle Modelling without Pretending to have to much a-priori Economic Theory," in *New Methods in Business Cycle Research*, ed. by C. Sims. Federal Reserve Bank of Minneapolis.

Shumway, R., and D. Stoffer (1982): "An approach to time series smoothing and forecasting using the EM algorithm," *Journal of Time Series Analysis*, 3, 253–264.

Stock, J. H., and M. W. Watson (2002a): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179.

——— (2002b): "Macroeconomic Forecasting Using Diffusion Indexes.," *Journal of Business and Economics Statistics*, 20, 147–162.

Watson, M. W., and R. F. Engle (1983): "Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models," *Journal of Econometrics*, 23, 385–400.

A Data Description

Number	Block	Name	Transf.	Availab.
		Benchmark dataset		
1		World market prices of raw materials in Euro. Index Total.	3	Jan-80
2	World	World market prices of raw materials in Euro. Index Total, excluding Energy	3	Jan-80
3	Prices	WORLD-MKT PRICES, ENERGY RAW MAT., CRUDE OIL	3	Jan-80
4		GOLD PRICE, US DOLLARS/FINE OUNCE,LONDON FIXING	3	Jan-80
5		Brent Crude-1 Month Fwd,fob US$/BBL converted in euro	3	May-85
6		trade with World (all entities), Export Value SA, not w.d. adj.	2	Jan-80
7	Trade	trade with World (all entities), Export Value, SA, not w.d. adj.	2	Jan-80
8		trade with World (all entities), Export Value, SA, not w.d. adj.	2	Jan-80
9		trade with World (all entities), Export Value, SA, not w.d. adj.	2	Jan-80
10		Retail trade, except of motor vehicles and motorcycles - W.d. and SA	2	Jan-80
11		Total industry - w.d. and SA	2	Jan-88
12		Total Industry (excluding construction) - W.d. and SA	2	Jan-85
13		Manufacturing - W.d. and SA	2	Jan-85
14		Construction - W.d. and SA	2	Jan-88
15		Total Industry excluding Construction and MIG Energy - W.d. and SA	2	Jan-85
16		Energy excluding NACE Rev.1 Section E - W.d. and SA	2	Jan-90
17		MIG Capital Goods Industry - W.d. and SA	2	Jan-80
18	Industrial	MIG Durable Consumer Goods Industry - W.d. and SA	2	Jan-90
19	Production	MIG Energy - Working day and SA	2	Jan-80
20		MIG Intermediate Goods Industry - Working day and SA	2	Jan-80
21		MIG Non-durable Consumer Goods Industry - W.d. and SA	2	Jan-85
22		Manufacture of basic metals - W.d. and SA	2	Jan-85
23		Manufacture of chemicals and chemical products - W.d. and SA	2	Jan-85
24		Manufacture of electrical machinery and apparatus n.e.c. - W.d. and SA	2	Jan-80
25		Manufacture of machinery and equipment n.e.c. - W.d. and SA	2	Jan-85
26		Manufacture of pulp, paper and paper products - W.d. and SA	2	Jan-85
27		Manufacture of rubber and plastic products - W.d. and SA	2	Jan-85
28		Industry Survey: Industrial Confidence Indicator - SA	1	Jan-85
29		Industry Survey: Production trend observed in recent months - SA	1	Jan-85
30		Industry Survey: Assessment of order-book levels - SA	1	Jan-85
31		Industry Survey: Assessment of export order-book levels - SA	1	Jan-85
32		Industry Survey: Assessment of stocks of finished products - SA	1	Jan-85
33	European	Industry Survey: Production expectations for the months ahead - SA	1	Jan-85
34	Commission	Industry Survey: Employment expectations for the months ahead - SA	-	Jan-85
35	Surveys	Consumer Survey: Consumer Confidence Indicator - SA	1	Jan-85
36		Consumer Survey: General economic situation over last 12 months - SA	1	Jan-85
37		Consumer Survey: General economic situation over next 12 months - SA	1	Jan-85
38		Consumer Survey: Unemployment expectations over next 12 months - SA	1	Jan-85
39		Construction Survey: Construction Confidence Indicator - SA	1	Jan-85
40		Construction Survey: Trend of activity compared with preceding months 1 SA	-	Jan-85
41		Construction Survey: Assessment of order books - SA	1	Jan-85
42		Construction Survey: Employment expectations for the months ahead - SA	1	Jan-85
43		Retail Trade Survey: Retail Confidence Indicator - SA	1	Jan-85
44		Retail Trade Survey: Present business situation - SA	1	Jan-85
45	Retail	Retail Trade Survey: Assessment of stocks - SA	1	Jan-85
46	Trade	Retail Trade Survey: Expected business situation - SA	1	Jan-85
47		Retail Trade Survey: Employment expectations - SA	1	Apr-85
48		New passenger car - W.d. and SA	2	Jan-90
49		Unemployment rate, Total (all ages), Total (male & female), SA, not w.d. adj	1	Jan-93
50		Index of Employment, Construction; SA, not w.d. adj	2	Jan-93
51	Labour	Index of Employment, Manufacturing; SA, not w.d. adj	2	Oct-89
52		Index of Employment, Total Industry; SA, not w.d. adj	2	Jan-93
53		Index of Employment, Total Industry (excluding construction);	2	Oct-89
54		EUROSTOXX 50 (RHS)	2	Dec-86
55	Financial	EUROSTOXX 325 (LHS)	2	Dec-86
56		S&P 500 COMPOSITE - PRICE INDEX	2	Jan-80
57		US, STOCK-EXCH. PRICES, DOW JONES, INDUSTRIAL AVERAGE, NSA	2	Jan-80
58		US, INT.RATE, MONEY-MKT, TREAS.BILLS, 3-MONTH, MKT YIELD	1	Jan-80
59		US, YIELD, SECOND.MKT, US TREASURY NOTES & BONDS, 10 YEARS	1	Jan-80
60		US, UNEMPLOYMENT RATE, SA	1	Jan-80
61	US	US, INDUST.PROD., TOTAL EXCL. CONSTRUCTION, SA	2	Jan-80
62		US, EMPLOYMENT, CIVILIAN, SA	2	Jan-80
63		US, RETAIL TRADE, VALUE (NAICS DEF.), SA	2	Jan-92
64		US, PRODUCTION EXPECTATIONS IN MANUF., NSA	1	Jan-80
65		US, CONSUMER EXPECTATIONS INDEX, NSA	1	Jan-80
66		10-year govt. bonds	1	Jan-80
67	Interest	Reuters.Money market.Euro.Euribor	1	Jan-80
68	Rates	1-year EURIBOR RATE	1	Jan-80
69		YIELD, SECOND MKT, GOVT BONDS, 2 YEARS	1	Jan-80
70		YIELD, SECOND MKT, GOVT BONDS, 5 YEARS	1	Jan-80
		Short Monthly dataset		
71		Manufacturing - employment	1	Aug-97
72		Manufacturing - new orders	1	Aug-97
73		Manufacturing - new export orders	1	Aug-97
74		Manufacturing - output	1	Aug-97
75		Manufacturing - purchasing manager index	1	Aug-97
76		Manufacturing - productivity	1	Jan-98
77	Purchasing	Manufacturing - quantity of purchases	1	Aug-97
78	Managers'	Manufacturing - supplier delivery times	1	Aug-97
79	Surveys	Manufacturing - stocks of finished goods	1	Aug-97
80		Manufacturing - stocks of purchases	1	Aug-97
81		Services - employment	1	Jul-98
82		Services - future business activity expectations	1	Jul-98
83		Services - new business	1	Jul-98
84		Services - outstanding business	1	Jul-98
85		Services - business activity	1	Jul-98
86		Services - productivity	1	Jan-98
		Quarterly dataset		

87		Total domestic- Level - Thousands of persons - SA	4	Q1-1980
88		Employees: total domestic- Level - Thousands of persons - NSA	4	Q1-1980
89	Employment:	Self-employed: total- Level - Thousands of persons - NSA	4	Q1-1980
90	by employment status,	Total industry- Level - Thousands of persons - NSA	4	Q1-1980
91	by economic activity	Construction- Level - Thousands of persons - NSA	4	Q1-1980
92		Trade and other - Level - Thousands of persons - NSA	4	Q1-1990
93		Total - index 2000 = 100 - SA	4	Q1-1995
94	Unit labour cost:	Industry, including energy - Index 2000 = 100 - SA	4	Q1-1995
95	by economic activity	Construction - Index 2000 = 100 - SA	4	Q1-1995
96		Trade and other - Index 2000 = 100 - SA	4	Q1-1995
97	Hourly labour cost	Whole economy excluding agriculture, fishing and government, SA	4	Q1-1996
98	US	UNIT LABOUR COSTS IN MANUFACTURING, SA	4	Q1-1959
99		GDP, AT MARKET PRICES - CHAINED 2000 USD SAAR	4	Q1-1959
100		Gross domestic product at market price - Chain linked - SA	4	Q1-1991
101		Final consumption of households and NPISH's - Chain linked - SA	4	Q1-1991
102	GDP: by expenditure	Final consumption of general government - Chain linked - SA	4	Q1-1991
103	components at	Gross fixed capital formation - Chain linked - SA	4	Q1-1991
104	constant prices	Exports of goods and services - Chain linked - SA	4	Q1-1991
105		Imports of goods and services - Chain linked - SA	4	Q1-1991
106		Gross value added at basic prices - SA	4	Q1-1995
107		Agricultural, hunting, forestry and fishing products - SA	4	Q1-1995
108	Value added:	Total industry - SA	4	Q1-1995
109	by economic activity	Construction - SA	4	Q1-1995
110	at constant prices	Trade and other - SA	4	Q1-1995
111		Financial intermediation, real estate - SA	4	Q1-1995
112		Other services - SA	4	Q1-1995
113		Taxes less subsidies on products - SA	4	Q1-1995
114	EC Survey	Industry Survey: Current level of capacity utilization	5	Q1-1980

Transformation code:
1: $y_{it} = (1 + L + L^2)(1 - L^3)Y_{it}$; 2: $y_{it} = (1 + L + L^2)(1 - L^3)\log(Y_{it})$;
3: $y_{it} = (1 + L + L^2)(1 - L^3)(1 - L^{12})\log(Y_{it})$; 4: $y_{it} = (1 - L^3)\log(Y_{it})$; 5: $y_{it} = (1 - L^3)(Y_{it})$

B Derivation of the EM iterations

Let us first derive the estimates for the case of $p = 1$. Let us recall that $\theta = (\Lambda, A_1, R, Q)$. For the model given by (4) the joint log-likelihood (for the observations and the latent factors) is given by:

$$\begin{aligned} l(Y,F;\theta) &= -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(f_0-\mu)'\Sigma^{-1}(f_0-\mu) \\ &- \frac{T}{2}\log|Q| - \frac{1}{2}\sum_{t=1}^{T}(f_t - A_1 f_{t-1})'Q^{-1}(f_t - A_1 f_{t-1}) \\ &- \frac{T}{2}\log|R| - \frac{1}{2}\sum_{t=1}^{T}(y_t - \Lambda f_t)'R^{-1}(y_t - \Lambda f_t) \\ &= -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(f_0-\mu)'\Sigma^{-1}(f_0-\mu) \\ &- \frac{T}{2}\log|Q| - \frac{1}{2}\mathrm{tr}\left[Q^{-1}\sum_{t=1}^{T}(f_t - A_1 f_{t-1})(f_t - A_1 f_{t-1})'\right] \\ &- \frac{T}{2}\log|R| - \frac{1}{2}\mathrm{tr}\left[R^{-1}\sum_{t=1}^{T}(y_t - \Lambda f_t)(y_t - \Lambda f_t)'\right] \end{aligned}$$

By differentiating $\mathrm{E}_{\theta(r)}\big[l(Y,F;\theta)|\Omega_T\big]$ with respect to A_1 we get

$$\begin{aligned} \frac{\partial \mathrm{E}_{\theta(r)}\big[l(Y,F;\theta)|\Omega_T\big]}{\partial A_1} &= -\frac{1}{2}\frac{\partial \mathrm{tr}\left\{Q^{-1}\sum_{t=1}^{T}\mathrm{E}_{\theta(r)}\big[(f_t - A_1 f_{t-1})(f_t - A_1 f_{t-1})'|\Omega_T\big]\right\}}{\partial A_1} \\ &= -Q^{-1}\sum_{t=1}^{T}\mathrm{E}_{\theta(r)}\big[f_t f_{t-1}'|\Omega_T\big] + Q^{-1}A_1\sum_{t=1}^{T}\mathrm{E}_{\theta(r)}\big[f_{t-1}f_{t-1}'|\Omega_T\big], \end{aligned}$$

and consequently

$$A_1(r+1) = \left(\sum_{t=1}^{T}\mathrm{E}_{\theta(r)}\big[f_t f_{t-1}'|\Omega_T\big]\right)\left(\sum_{t=1}^{T}\mathrm{E}_{\theta(r)}\big[f_{t-1}f_{t-1}'|\Omega_T\big]\right)^{-1},$$

as provided in the main text. Similarly we can obtain $\Lambda(r+1)$, $Q(r+1)$, and $R(r+1)$, given by the formulas (7)-(9)

In case y_t contains missing values, (10) no longer holds and the formulas for $\Lambda(r+1)$ and $R(r+1)$ need to be modified. Let us differentiate $\mathrm{E}_{\theta(r)}\big[l(Y,F;\theta)|\Omega_T\big]$ with respect to Λ:

$$\frac{\partial \mathrm{E}_{\theta(r)}\big[l(Y,F;\theta)|\Omega_T\big]}{\partial \Lambda} = -\frac{1}{2}\frac{\partial \mathrm{tr}\Big\{R^{-1}\sum_{t=1}^{T}\mathrm{E}_{\theta(r)}\big[(y_t - \Lambda f_t)(y_t - \Lambda f_t)'|\Omega_T\big]\Big\}}{\partial \Lambda}$$

and let us have a closer look at $\mathrm{E}\big[(y_t - \Lambda f_t)(y_t - \Lambda f_t)'|\Omega_T\big]$ (to simplify the notation we skip the subscript $\theta(r)$).

Let

$$y_t = W_t y_t + (I - W_t)y_t = y_t^{(1)} + y_t^{(2)},$$

where W_t is a diagonal matrix with ones corresponding to the non-missing entries in y_t and 0 otherwise. ($y_t^{(1)}$ contains the non-missing observations at time t with 0 in place of the missing ones.)

We have:

$$\begin{aligned}(y_t-\Lambda f_t)(y_t-\Lambda f_t)' &= \Big(W_t(y_t-\Lambda f_t)+(I-W_t)(y_t-\Lambda f_t)\Big)\Big(W_t(y_t-\Lambda f_t)+(I-W_t)(y_t-\Lambda f_t)\Big)'\\ &= W_t(y_t-\Lambda f_t)(y_t-\Lambda f_t)'W_t+(I-W_t)(y_t-\Lambda f_t)(y_t-\Lambda f_t)'(I-W_t)\\ &+ W_t(y_t-\Lambda f_t)(y_t-\Lambda f_t)'(I-W_t)+(I-W_t)(y_t-\Lambda f_t)(y_t-\Lambda f_t)'W_t\,.\end{aligned}$$

By the law of iterated expectations:

$$\mathrm{E}\big[(y_t-\Lambda f_t)(y_t-\Lambda f_t)'|\Omega_T\big]=\mathrm{E}\Big[\mathrm{E}\big[(y_t-\Lambda f_t)(y_t-\Lambda f_t)'|F,\Omega_T\big]|\Omega_T\Big]\,.$$

As

$$\begin{aligned}\mathrm{E}\big[W_t(y_t-\Lambda f_t)(y_t-\Lambda f_t)'(I-W_t)|F,\Omega_T\big] &= 0\,,\\ \mathrm{E}\big[(I-W_t)(y_t-\Lambda f_t)(y_t-\Lambda f_t)'(I-W_t)|F,\Omega_T\big] &= (I-W_t)R(r)(I-W_t)\end{aligned}$$

and

$$\begin{aligned}&\mathrm{E}\big[W_t(y_t-\Lambda f_t)(y_t-\Lambda f_t)'W_t|\Omega_T\big]\\ &\quad=W_ty_ty_t'W_t-W_ty_t\mathrm{E}\big[f_t'|\Omega_T\big]\Lambda'W_t-W_t\Lambda\mathrm{E}\big[f_t|\Omega_T\big]y_t'W_t+W_t\Lambda\mathrm{E}\big[f_tf_t'|\Omega_T\big]\Lambda'W_t\\ &\quad=y_t^{(1)}y_t^{(1)\prime}-y_t^{(1)}\mathrm{E}\big[f_t'|\Omega_T\big]\Lambda'W_t-W_t\Lambda\mathrm{E}\big[f_t|\Omega_T\big]y_t^{(1)\prime}+W_t\Lambda\mathrm{E}\big[f_tf_t'|\Omega_T\big]\Lambda'W_t\,,\end{aligned}$$

we get:

$$\begin{aligned}&\mathrm{E}\big[(y_t-\Lambda f_t)(y_t-\Lambda f_t)'|\Omega_T\big]=\\ &y_t^{(1)}y_t^{(1)\prime}-y_t^{(1)}\mathrm{E}\big[f_t'|\Omega_T\big]\Lambda'W_t-W_t\Lambda\mathrm{E}\big[f_t|\Omega_T\big]y_t^{(1)\prime}+W_t\Lambda\mathrm{E}\big[f_tf_t'|\Omega_T\big]\Lambda'W_t+(I-W_t)R(r)(I-W_t)\,.\end{aligned}$$

Consequently:

$$\begin{aligned}\frac{\partial\mathrm{tr}\Big\{R^{-1}\mathrm{E}\big[(y_t-\Lambda f_t)(y_t-\Lambda f_t)'|\Omega_T\big]\Big\}}{\partial\Lambda} &= -2W_tR^{-1}y_t^{(1)}\mathrm{E}\big[f_t'|\Omega_T\big]+2W_tR^{-1}W_t\Lambda\mathrm{E}\big[f_tf_t'|\Omega_T\big]\\ &= -2R^{-1}y_t^{(1)}\mathrm{E}\big[f_t'|\Omega_T\big]+2R^{-1}W_t\Lambda\mathrm{E}\big[f_tf_t'|\Omega_T\big]\,.\end{aligned}$$

From

$$\sum_{t=1}^{T}\frac{\partial\mathrm{tr}\Big\{R^{-1}\mathrm{E}_{\theta(r)}\big[(y_t-\Lambda f_t)(y_t-\Lambda f_t)'|\Omega_T\big]\Big\}}{\partial\Lambda}\bigg|_{\Lambda=\Lambda(r+1)}=0$$

follows

$$\sum_{t=1}^{T}y_t^{(1)}\mathrm{E}_{\theta(r)}\big[f_t'|\Omega_T\big]=\sum_{t=1}^{T}W_t\Lambda(r+1)\mathrm{E}_{\theta(r)}\big[f_tf_t'|\Omega_T\big]\,.$$

Equivalently (as $\mathrm{vec}(ABC)=(C'\otimes A)\mathrm{vec}(B)$) we have

$$\mathrm{vec}\left(\sum_{t=1}^{T}y_t^{(1)}\mathrm{E}_{\theta(r)}\big[f_t'|\Omega_T\big]\right)=\left(\sum_{t=1}^{T}\mathrm{E}_{\theta(r)}\big[f_tf_t'|\Omega_T\big]\otimes W_t\right)\mathrm{vec}\big(\Lambda(r+1)\big)$$

hence

$$\operatorname{vec}\big(\Lambda(r+1)\big) = \left(\sum_{t=1}^{T} \mathrm{E}_{\theta(r)}\big[f_t f_t'|\Omega_T\big] \otimes W_t\right)^{-1} \operatorname{vec}\left(\sum_{t=1}^{T} y_t^{(1)} \mathrm{E}_{\theta(r)}\big[f_t'|\Omega_T\big]\right).$$

In the similar fashion we obtain

$$\begin{aligned} R(r+1) &= \operatorname{diag}\Bigg(\frac{1}{T}\sum_{t=1}^{T}\Big(y_t^{(1)}y_t^{(1)\prime} - y_t^{(1)}\mathrm{E}_{\theta(r)}\big[f_t'|\Omega_T\big]\Lambda(r+1)'W_t - W_t\Lambda(r+1)\mathrm{E}_{\theta(r)}\big[f_t|\Omega_T\big]y_t^{(1)\prime} \\ &+ W_t\Lambda(r+1)\mathrm{E}_{\theta(r)}\big[f_t f_t'|\Omega_T\big]\Lambda(r+1)'W_t + (I-W_t)R(r)(I-W_t)\Big)\Bigg). \end{aligned}$$

Let us now consider the case of $p > 1$. We can write the log-likelihood:

$$\begin{aligned} l(Y,F;\theta) &= -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(\tilde{f}_0-\mu)'\Sigma^{-1}(\tilde{f}_0-\mu) \\ &- \frac{r}{2}\log|Q| - \frac{1}{2}\mathrm{tr}\left[Q^{-1}\sum_{t=1}^{T}(f_t - A\tilde{f}_{t-1})(f_t - A\tilde{f}_{t-1})'\right] \\ &- \frac{n}{2}\log|R| - \frac{1}{2}\mathrm{tr}\left[R^{-1}\sum_{t=1}^{T}(y_t - \Lambda f_t)(y_t - \Lambda f_t)'\right], \end{aligned}$$

where $\tilde{f}_{t-1} = [f_{t-1}', \ldots, f_{t-p}']'$ and $A = [A_1, \ldots, A_p]$.

Consequently (6) and (8) should be modified as:

$$A(r+1) = \left(\sum_{t=1}^{T}\mathrm{E}_{\theta(r)}\big[f_t\tilde{f}_{t-1}'|\Omega_T\big]\right)\left(\sum_{t=1}^{T}\mathrm{E}_{\theta(r)}\big[\tilde{f}_{t-1}\tilde{f}_{t-1}'|\Omega_T\big]\right)^{-1}$$

and

$$Q(r+1) = \frac{1}{T}\left(\sum_{t=1}^{T}\mathrm{E}_{\theta(r)}\big[f_t f_t'|\Omega_T\big] - A(r+1)\sum_{t=1}^{T}\mathrm{E}_{\theta(r)}\big[\tilde{f}_{t-1}f_t'|\Omega_T\big]\right).$$

The conditional moments of the factors $\mathrm{E}_{\theta(r)}\big[f_t\tilde{f}_{t-1}'|\Omega_T\big]$, $\mathrm{E}_{\theta(r)}\big[\tilde{f}_{t-1}\tilde{f}_{t-1}'|\Omega_T\big]$, $\mathrm{E}_{\theta(r)}\big[f_t f_t'|\Omega_T\big]$ can be obtained by running the Kalman filter on the following state space form:

$$Y_t = \begin{bmatrix} \Lambda & 0 & \ldots & 0 \end{bmatrix}\begin{bmatrix} f_t \\ f_{t-1} \\ \vdots \\ f_{t-p+1} \end{bmatrix} + \epsilon_t \qquad \epsilon_t \sim N(0,R),$$

$$\begin{bmatrix} f_t \\ f_{t-1} \\ \vdots \\ f_{t-p+1} \end{bmatrix} = \begin{bmatrix} A_1 & A_2 & \cdots & A_p \\ I & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & I & 0 \end{bmatrix}\begin{bmatrix} f_{t-1} \\ f_{t-2} \\ \vdots \\ f_{t-p} \end{bmatrix} + u_t \qquad u_t \sim N\left(0, \begin{bmatrix} Q & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix}\right).$$

C Convergence checks

We perform simulations similar to those in Doz, Giannone, and Reichlin (2006). More precisely we simulate the data from the following factor model:

$$\begin{aligned} y_t &= \Lambda f_t + \xi_t\,, \\ f_t &= A f_{t-1} + u_t \qquad u_t \sim N(0, I_r)\,, \\ \xi_t &= D\xi_{t-1} + v_t \qquad v_t \sim N(0, \Phi)\,, \end{aligned}$$

where

$$\begin{aligned} &\Lambda_{ij} \quad \text{i.i.d.} \quad N(0,1)\,, \\ &A_{ij} \quad = \quad \begin{cases} \rho, & i=j \\ 0, & i\neq j \end{cases}\,, \qquad D_{ij} = \begin{cases} d, & i=j \\ 0, & i\neq j \end{cases}\,, \\ &\Phi \quad = \quad \tau^{|i-j|}(1-d^2)\sqrt{\alpha_i\alpha_j}\,, \quad \alpha_i = \frac{\beta_i}{1-\beta_i}\frac{1}{1-\rho^2}\sum_{j=1}^{r}\Lambda_{ij}^2\,, \quad \beta_i \text{ i.i.d. } U\big([u,1-u]\big)\,, \end{aligned}$$

see Stock and Watson (2002a) or Doz, Giannone, and Reichlin (2006) for the discussion. This model allows for serial and (some degree of) cross-correlation in the idiosyncratic components.

The data are generated with $r = 3$, $\rho = 0.9$, $d = 0.5$, $u = 0.1$ and $\tau = 0.5$. Subsequently, we set $k\%$ of the data as missing and we estimate the parameters and the factors. We use the same stopping criterion as Doz, Giannone, and Reichlin (2006).

To assess the precision of the estimates of the factors we follow Stock and Watson (2002a) and Doz, Giannone, and Reichlin (2006) and use the trace R^2 of the regression of the estimated factors on the true ones:

$$\frac{\text{Trace}\big(F'\hat{F}\left(\hat{F}'\hat{F}\right)^{-1}\hat{F}'F\big)}{\text{Trace}\big(F'F\big)}\,.$$

This measure is smaller than 1 and tends to 1 with the increasing canonical correlation between the estimated and the true factors.

Table 5 presents the trace statistics for different cross-section size n, different series length T and different missing-data ratio k.

We can see that in each case the factors converge to the true ones with increasing T and n albeit for larger share of missing data the convergence is slightly slower.

In the case with missing data the convergence of the EM algorithm is slower as well. To speed up the computations, the solutions proposed in Koopman and Durbin (2000) and Jungbacker and Koopman (2008) could be considered.[17]

D Computation of the news

As in the Section 2.3 let Ω_{v-1} and Ω_v be two consecutive vintages of data and let I_v be the *news* content of Ω_v orthogonal to Ω_{v-1}. We will derive the formula for the projection $\mathbb{P}\left(y_{j,t_j}|I_v\right)$ for a general form of I_v.

[17] In the case of the former the diagonal form of the covariance matrix of the idiosyncratic component can be exploited, in the case of the latter, the fact that the size of observation vector is large relative to the size of the state vector.

Table 5: Trace statistics for the factor estimates

T/n	10	25	50	100
No missing data				
50	0.54	0.64	0.69	0.72
100	0.61	0.76	0.81	0.83
10% missing data				
50	0.53	0.63	0.69	0.72
100	0.61	0.76	0.81	0.83
20% missing data				
50	0.53	0.63	0.68	0.71
100	0.60	0.75	0.80	0.83
40% missing data				
50	0.50	0.60	0.67	0.70
100	0.59	0.73	0.79	0.82

Notes: Table reports trace statistics for the factor estimates for different ratios of missing data for data simulated from a factor model. T and n refer to the sample and cross-section size respectively.

Without a loss of generality, let us assume that Ω_v contains new releases for the first K variables for the reference periods t_1, ..., t_K, respectively. Then we have

$$I_v = \begin{bmatrix} y_{1,t_1} - \mathbb{P}\left(y_{1,t_1}|\Omega_{v-1}\right) \\ y_{2,t_2} - \mathbb{P}\left(y_{2,t_2}|\Omega_{v-1}\right) \\ \vdots \\ y_{K,t_K} - \mathbb{P}\left(y_{K,t_K}|\Omega_{v-1}\right) \end{bmatrix}.$$

The projection is given by

$$\mathbb{P}\left(y_{j,t_j}|I_v\right) = \mathrm{E}\left(y_{j,t_j} I_v'\right) \mathrm{E}\left(I_v I_v'\right)^{-1} I_v \,, \tag{12}$$

where

$$\mathrm{E}\left(y_{j,t_j} I_v'\right) = \begin{bmatrix} \mathrm{E}\Big(y_{j,t_j}\left(y_{1,t_1} - \mathbb{P}\left(y_{1,t_1}|\Omega_{v-1}\right)\right)\Big) \\ \mathrm{E}\Big(y_{j,t_j}\left(y_{2,t_2} - \mathbb{P}\left(y_{2,t_2}|\Omega_{v-1}\right)\right)\Big) \\ \vdots \\ \mathrm{E}\Big(y_{j,t_j}\left(y_{K,t_K} - \mathbb{P}\left(y_{K,t_K}|\Omega_{v-1}\right)\right)\Big) \end{bmatrix}'$$

and

$$\mathrm{E}\left(I_v I_v'\right) = \Big[\ \mathrm{E}\Big(\big(y_{i,t_i} - \mathbb{P}\left(y_{i,t_i}|\Omega_{v-1}\right)\big)\big(y_{k,t_k} - \mathbb{P}\left(y_{k,t_k}|\Omega_{v-1}\right)\big)\Big)\ \Big]_{\{i=1,\ldots,K;k=1,\ldots,K\}}.$$

In order to obtain (12) we need to calculate $\mathrm{E}\Big(y_{j,t_j}\left(y_{i,t_i} - \mathbb{P}\left(y_{i,t_i}|\Omega_{v-1}\right)\right)\Big)$ and $\mathrm{E}\Big(\big(y_{i,t_i} - \mathbb{P}\left(y_{i,t_i}|\Omega_{v-1}\right)\big)\big(y_{k,t_k} - \mathbb{P}\left(y_{k,t_k}|\Omega_{v-1}\right)\big)\Big)$.

Given the model (4) we can write

$$\begin{aligned} y_{j,t_j} &= \lambda_j f_{t_j} + \xi_{j,t_j} \,, \\ y_{j,t_j} - y_{j,t_j|\Omega_{v-1}} &= \lambda_j \left(f_{t_j} - f_{t_j|\Omega_{v-1}}\right) + \xi_{j,t_j} \,. \end{aligned}$$

Therefore we have

$$\begin{aligned}
\mathrm{E}\Big(y_{j,t_j}\left(y_{i,t_i}-y_{i,t_i|\Omega_{v-1}}\right)\Big) &= \lambda_j\mathrm{E}\Big(f_{t_j}\left(f_{t_i}-f_{t_i|\Omega_{v-1}}\right)'\Big)\lambda_i'+\mathrm{E}\Big(\xi_{j,t_j}\left(f_{t_i}-f_{t_i|\Omega_{v-1}}\right)'\Big)\lambda_i' \\
&= \lambda_j\mathrm{E}\Big(\left(f_{t_j}-f_{t_j|\Omega_{v-1}}\right)\left(f_{t_i}-f_{t_i|\Omega_{v-1}}\right)'\Big)\lambda_i'+\lambda_j\mathrm{E}\Big(f_{t_j|\Omega_{v-1}}\left(f_{t_i}-f_{t_i|\Omega_{v-1}}\right)'\Big)\lambda_i' \\
&= \lambda_j\mathrm{E}\Big(\left(f_{t_j}-f_{t_j|\Omega_{v-1}}\right)\left(f_{t_i}-f_{t_i|\Omega_{v-1}}\right)'\Big)\lambda_i'
\end{aligned}$$

and

$$\mathrm{E}\Big(\left(y_{i,t_i}-y_{i,t_i|\Omega_{v-1}}\right)\left(y_{k,t_k}-y_{k,t_k|\Omega_{v-1}}\right)'\Big) = \lambda_i\mathrm{E}\Big(\left(f_{t_i}-f_{t_i|\Omega_{v-1}}\right)\left(f_{t_k}-f_{t_k|\Omega_{v-1}}\right)'\Big)\lambda_k'+\mathrm{E}\Big(\xi_{i,t_i}\xi_{k,t_k}'\Big).$$

In the case that $t_j = t_i$ the expectation $\mathrm{E}\Big(\left(f_{t_j}-f_{t_j|\Omega_{v-1}}\right)\left(f_{t_i}-f_{t_i|\Omega_{v-1}}\right)'\Big)$ is returned by the Kalman smoother. To obtain the expectations for $t_j \neq t_i$ one can augment the vector of states by appropriate number of their lags.