Estimation of Common Factors under Cross-Sectional and Temporal Aggregation Constraints: Nowcasting Monthly GDP and its Main Components

Proietti, Tommaso

SEFEMEQ, Faculty of Economics, University of Rome "Tor Vergata"

22 January 2008

Estimation of Common Factors under Cross-Sectional and Temporal Aggregation Constraints: Nowcasting Monthly GDP and its Main Components

Tommaso Proietti*
SEFEMEQ, University of Rome "Tor Vergata"

Abstract

The paper estimates a large-scale mixed-frequency dynamic factor model for the euro area, using monthly series along with Gross Domestic Product (GDP) and its main components, obtained from the quarterly national accounts. The latter define broad measures of real economic activity (such as GDP and its decomposition by expenditure type and by branch of activity) that we are willing to include in the factor model, in order to improve its coverage of the economy and thus the representativeness of the factors. The main problem with their inclusion is not one of model consistency, but rather of data availability and timeliness, as the national accounts series are quarterly and are available with a large publication lag. Our model is a traditional dynamic factor model formulated at the monthly frequency in terms of the stationary representation of the variables, which however becomes nonlinear when the observational constraints are taken into account. These are of two kinds: nonlinear temporal aggregation constraints, due to the fact that the model is formulated in terms of the unobserved monthly logarithmic changes, but we observe only the sum of the monthly levels within a quarter, and nonlinear cross-sectional constraints, since GDP and its main components are linked by the national accounts identities, but the series are expressed in chained volumes. The paper provides an exact treatment of the observational constraints and proposes iterative algorithms for estimating the parameters of the factor model and for signal extraction, thereby producing nowcasts of monthly gross domestic product and its main components, as well as measures of their reliability.

Keywords: Dynamic Factor Models; EM algorithm; Non Linear State Space Models; Temporal Disaggregation; Nonlinear Smoothing; Monthly GDP; Chain-linking.

*Address for Correspondence: SEFEMEQ, Facoltà di Economia, Università di Roma "Tor Vergata", Via Columbia 2, 00133 Roma, Italy. E-mail tommaso.proietti@uniroma2.it

1 Introduction

Large scale factor models aim at extracting the main economic signals from a very large number of time series. The underlying idea is that the comovements among economic time series can be traced to a limited number of common factors. Factor models have been used in an increasing number of applications. The two most prominent areas are the construction of synthetic indicators, such as coincident indicators of real economic activity (Forni et al., 2000, 2001) and core inflation (Cristadoro et al., 2005), and forecasting macroeconomic variables (Stock and Watson, 2002a, Forni et al. 2005), in which case the information contained in large number of economic indicators is summarized in a few latent factors, which are then employed to forecast real output growth, or inflation. Other areas of applications are surveyed in Stock and Watson (2006).

On the other hand, national accounts (NA) statistics provide a comprehensive and detailed record of the economic activities taking place within an economy, which are translated into a set of coherent and integrated measures of economic activity. The most comprehensive measure is provided by Gross Domestic Product (GDP); furthermore, the aggregates that arise from its decomposition according to the expenditure and the output approach (e.g. final consumption, gross capital formation, sectorial value added) are among the most relevant economic statistics for purposes of macroeconomic analysis and policy-making.

Hence, the NA aggregates can be considered as aggregate indicators of economic activity based on a set of definitions, concepts, classifications and accounting rules that are internationally agreed. The main problem is their observation frequency, which at present is quarterly for the euro area, and their timeliness, i.e., the fact that they are made available with considerable delay. A related point is that they are first released as preliminary estimates and then revised as new information accrues. Their lack of timeliness is a direct consequence of their comprehensiveness and generality: their estimation requires a lot of information from the institutional units; a large part of this information is used to construct the monthly time series that are typically considered by factor models (with some exceptions, e.g. business and consumer survey).

The aim of this paper is to estimate a large scale factor model of the euro area economy which combines the monthly information carried by a number of economic indicators (concerning industrial production, construction, retail sales, financial intermediation, employment and wages, exchange rates, external trade and business and consumer surveys) with the quarterly national accounts series. In particular, we consider a panel of 149 series, referring to the euro area for the period from January 1995 to June 2007, 17 of which are NA series and concern quarterly real GDP and its breakdown according to the expenditure and the output approaches. The presence of these series raises the fundamental issue of incorporating the observational constraints into the estimation process. The issue has two facets: one of temporal aggregation and one of contemporaneous aggregation. As far as the former is concerned, the factor model is specified in terms of the stationary representation of the series; our series can be taken to be stationary in terms of the logarithmic change with respect to the previous month (assuming that all are nonseasonal or seasonally adjusted). For the NA series the monthly changes are unobserved. What we observe are the quarterly totals, i.e. the sum of the levels of the three months making up the quarter. This simple fact renders the observational constraint nonlinear. Secondly, the NA series are subject to accounting identities that, due to chain linking, hold when the data are expressed at the prices of the previous year (see Eurostat, 1999, Bloem, et. al, 2001). This again makes the cross-sectional constraints nonlinear.

The introduction of the NA series in the model can be considered as the main contribution of this paper. Their consideration is essential to improve the coverage of the economy and the representativeness of the factors. The main problem with their inclusion is not one of model consistency, but rather of data availability and timeliness; as a matter of facts, it allows to incorporate in the factor estimates the information arising from core measures such as GDP, final consumption expenditures, sectorial value added and other main NA aggregates. The inclusion entails that contemporaneous aggregation constraints arising from national accounts identities are taken into consideration. Secondly, as a by product our model produces nowcasts of monthly GDP and its components, along with measures of their reliability. Not only the factor estimates will benefit from the inclusion of GDP and its components, but also the disaggregate estimates of GDP will embody a large information set.

The availability of an indicator of monthly GDP is an important addition to the set of available economic statistics. A variety of approaches are available, ranging from linear univariate temporal disaggregation using the factors as monthly indicators, to multivariate parametric models, and a relatively large literature is already available on this or related topics. See, among others, Angelini, Henry and Marcellino (2004), Mariano and Murasawa (2003), Mönch, E., and Uhlig, H. (2004), Proietti and Moauro (2006), Giannone, Reichlin and Small (2006), Breitung and Schumacher (2006), Bańbura and Ruenstler (2007), Aruoba, Diebold and Scotti (2007), Altissimo et al. (2007).

Our contribution to the literature is to provide the joint temporal disaggregation of the NA series, within a large scale factor model, whose specification includes the NA series themselves, giving an exact treatment of the temporal and cross-sectional aggregation constraints. The temporal aggregation constraints are enforced by an iterative nonlinear smoothing algorithm. The cross-sectional constraints are enforced by a multistep procedure that de-chains the estimated monthly values, expressing them at the average prices of the previous year, and projects the estimates on the subspace of the constraints. The dechaining procedure is in line with that advocated by the IMF manual (see Bloem *et al.*, 2001). Finally, the series are chained back and expressed in volumes at the prices of the reference year.

As a result the monthly estimates of the NA series are consistent with the temporal aggregation constraints (the quarterly sums are equal to the data released by Eurostat) and the accounting identities, when the series are expressed at the prices of the previous year. Another advantage of our approach is the possibility to assess the reliability of the monthly GDP estimates.

The estimation of the factor models is carried out by an iterative procedure. Each iteration consists of two step. Given the availability of a preliminary estimate of the monthly NA series, the first step estimates the parameters using the EM algorithm or principal component analysis. Conditional on the parameter estimates, the second step obtains the estimates of the factors and the disaggregate NA series by solving a nonlinear smoothing problem.

The paper is structured as follows: section 2 provides a description of the panel of time series available. In section 3 we discuss the specification of the linear dynamic factor model for the complete monthly dataset, that is assuming that the panel time series were balanced and characterised by the same observation frequency. Estimation of the model parameters by the EM algorithm and by principal components is discussed in sections 4.1 and 4.2. We then discuss the implications of temporal aggregation in section 5. The constraints are enforced by the nonlinear smoothing algorithm described in section 6, in which we discuss the modified state space model that arises and its sequential constrained estimation of the factors and the missing monthly values. Section 7 deals with the statistical treatment of the cross-sectional constraints that arise from the accounting identities. The main estimation results for the euro area are presented and discussed in section 8. Finally, we draw our conclusion and hint at some future developments (section 9).

2 Description of the dataset

The available data consist of 132 monthly and 17 quarterly time series (i.e. a total of 149 series) for the period starting in January (1st quarter of) 1995 and ending in June (second quarter) of 2006, for a total of 150 monthly observation (38 quarterly observations). The series, extracted from the Europa database (http://epp.eurostat.ec.europa.eu/), are listed in Appendix 1 and can be grouped under the following main headings.

National accounts: 17 quarterly time series concerning the euro area GDP and its main components, the breakdown of total GDP by the output the expenditure approaches. The complete list is provided in table 1. All the series are expressed in millions of euro, chain-linked volumes, reference year 2000. When expressed at the prices of the previous year (as it occurs for the values of the year 2001, which are expressed in 2000 euros) the series are subject to contemporaneous aggregation constraints. The role of these constraints for the estimation of the disaggregate time series will be the topic of section 7.

Industry: 53 monthly time series. Index of industrial production (25 series); Monthly turnover index (7 series); Monthly indices of new orders (6 series); Volume of work done (hours worked) (8 series); Gross wages and salaries (7 series); see Table 2.

Construction: 7 monthly time series. Monthly production index (3 series); Monthly indices of labour input (3 series); Building permits (1 series); see Table 3.

Retail Trade: 28 monthly time series. Index of turnover (13 series); Index of deflated turnover (13 series); Employment (1 series); Car registration (1 series); see Table 3.

Monetary and Financial indicators: 13 monthly time series. Exchange rates (6 series); Money supply (3 series); Share price index (1 series); Interest rates (3 series); see Table 4.

Labour market: 5 monthly time series. Harmonised unemployment rates (5 series); see Table 4

External trade 4 monthly time series. Total imports and exports, trade value and volume index; see Table 4.

Business and consumer surveys: 22 monthly time series. Industry (5 time series); Construction (5 time series); Retail sale (7 series); Consumer surveys (6 series); see Table 5.

All the series are seasonally adjusted and refer to the euro area with 12 member states (i.e. the European Monetary Union excluding Slovenia, Cyprus and Malta). Only the Business and Consumer surveys, produced by DG ECFIN, refer to the euro area with 13 member states (including Slovenia). The choice was made necessary by data availability. The set of series can be considered as a unbalanced sample of the euro area economy which tends to over-represent the industrial sector. As it is well known, the service sector is under represented in the short run economic indicators. This is why we think that including the national accounts redresses the balance and improves upon the coverage of the factor model.

3 The Complete Data Factor Model

This section discusses the specification of the dynamic factor model for a balanced panel of time series characterized by the same observation frequency. The issue of temporal aggregation will be deferred to section 5. Thus, let us suppose that the N time series are fully available and let us denote the individual time series in the original scale of measurement by $Y_{it}, i = 1, \ldots, N, t = 0, 1, \ldots, n$. We also assume that the series can be rendered stationary by the transformation $y_{it} - \varphi_i y_{i,t-1}, t = 1, \ldots, n$, where y_{it} is the Box-Cox transformation (Box and Cox, 1964) with parameter λ_i of the original series,

$$y_{it} = \begin{cases} \frac{Y_{it}^{\lambda_i} - 1}{\lambda_i}, & \lambda_i \neq 0, \\ \ln Y_{it}, & \lambda_i = 0, \end{cases}$$

and $\varphi_i = 1$ if the series is difference stationary and 0 otherwise. For the series considered in our application, we can assume that the monthly logarithmic changes are stationary, so that $\lambda_i = 1$ and $\varphi_i = 1$, except for the Business and Consumer Survey series, for which $\lambda_i = 0$ and $\varphi_i = 0$.

The factor model that we formulate for the complete monthly series (i.e., the model that would be entertained if a complete set of N monthly time series were available) is a standard dynamic factor model, according to which the series are conditionally independent, given a set of common factors. The common factors are generated by a stationary first order vector autoregressive process. The model for the i-th time series is formulated as follows:

$$\begin{array}{rcl} y_{it} & = & \varphi_i y_{i,t-1} + \mu_i + \sigma_i x_{it}, \quad i = 1, \ldots, N, t = 1, \ldots, n, \\ x_{it} & = & \boldsymbol{\theta}_i' \mathbf{f}_t + \xi_{it}, \quad \xi_{it} \sim \mathrm{NID}(0, \psi_i), \\ \mathbf{f}_t & = & \boldsymbol{\Phi} \mathbf{f}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \mathrm{NID}(\mathbf{0}, \boldsymbol{\Sigma}_\eta); \end{array} \tag{1}$$

here μ_i represents the mean of the stationary transformation $y_{it} - \varphi_i y_{i,t-1}$, σ_i is its standard deviation, and x_{it} is the standardized stationary transformation of the original time series. The latter is expressed as a linear combination of K stationary common factors, $\mathbf{f}_t$, with zero mean, with weights collected in the $K \times 1$ vector $\boldsymbol{\theta}_i$ (factor loadings), plus an idiosyncratic component, ξ_{it}. The idiosyncratic component is orthogonal to the factors.

If we further let $\Delta y_{it} = y_{it} - \varphi_i y_{i,t-1}$ and $\Delta \mathbf{y}_t$ denote the stack of the stationary series, $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_N]'$, $\mathbf{D} = \mathrm{diag}(\sigma_1, \ldots, \sigma_N)$, and similarly $\mathbf{x}_t = [x_{1t}, \ldots, x_{Nt}]'$, we can write $\Delta \mathbf{y}_t = \boldsymbol{\mu} + \mathbf{D}\mathbf{x}_t$, and the model for $\mathbf{x}_t$ has state space representation:

$$\begin{array}{rcl} \mathbf{x}_t & = & \boldsymbol{\Theta} \mathbf{f}_t + \boldsymbol{\xi}_t, \quad \boldsymbol{\xi}_t \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi}) \\ \mathbf{f}_t & = & \boldsymbol{\Phi} \mathbf{f}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \mathrm{NID}(\mathbf{0}, \boldsymbol{\Sigma}_\eta) \end{array} \tag{2}$$

where $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N]'$ and $\boldsymbol{\Psi} = \mathrm{diag}\{\psi_1, \ldots, \psi_N\}$, $\mathrm{E}(\boldsymbol{\xi}_t \boldsymbol{\eta}_t') = \mathbf{0}$ and $\mathbf{f}_0 \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_f)$, where $\boldsymbol{\Sigma}_f$ satisfies the matrix equation $\boldsymbol{\Sigma}_f = \boldsymbol{\Phi} \boldsymbol{\Sigma}_f \boldsymbol{\Phi}' + \boldsymbol{\Sigma}_\eta$.

The model needs not be interpreted as a strict factor model, in the sense that we can relax to a certain extent the assumption of uncorrelatedness of the idiosyncratic component, allowing for serial and cross-sectional correlation for the idiosyncratic component. Overall, we can allow $\mathbf{x}_t$ to have an approximate factor structure in the sense specified by Bai (2003), or by Forni et al. (2005).

As it is well known, the factor model is identified up to an invertible $K \times K$ matrix. A unique solution is obtained by imposing K^2 restrictions. We identify our factor model using the restriction that the upper $K \times K$ block of the loadings matrix is equal to the identity matrix, that is $\boldsymbol{\Theta} = [\mathbf{I}_K, \boldsymbol{\Theta}^{*'}]'$. The restriction exactly identifies the model; see Geweke and Singleton (1981), proposition 2.

Let us define the parameter vector $\boldsymbol{\Xi} = [\mathrm{vec}(\boldsymbol{\Theta}^*)', \mathrm{vec}(\boldsymbol{\Phi})', \mathrm{vech}(\boldsymbol{\Sigma}_\eta), \psi_1, \ldots, \psi_N]'$. For small N the parameters can be estimated by maximum likelihood, where the likelihood is evaluated by the Kalman filter (KF) via the prediction error decomposition, using a numerical quasi-Newton method. An application is

Stock and Watson (1991). With large N, the evaluation of the likelihood is still efficiently performed by the KF; however the difficulty with maximising the likelihood via gradiend based methods is due to the high dimensionality of Ξ, which has $NK+N+K^2$ unrestricted elements. In our application, in which $N = 149$ and $K = 6$, the number of unrestricted parameters is 1079.

A computationally viable alternative is to use the Expectation- Maximization (EM) algorithm of Dempster et al. (1977). The EM algorithm for state space models was introduced by Shumway and Stoffer (1982) and Watson and Engle (1983). For N large, an alternative asymptotically equivalent estimation strategy is to use principal components analysis, when we allow the number of time series N, or both N and n, to go to infinity. In the next section we review the two estimation strategies in some detail.

4 Estimation of the Complete Data Factor Model

In this section we provide the details concerning the estimation of the stationary dynamic factor model (2), under the assumption that the N standardized time series collected in the vector $\mathbf{x}_t$ are fully observed, for $t = 1, \ldots, n$. We assume that the number of factors is know, or it has been estimated according to the information criteria proposed by Bai and Ng (2002), and extended by Amenegual and Watson (2007) to a dynamic setting.

4.1 The EM Algorithm

The derivation of the EM algorithm made in this section is based on Shumway and Stoffer (1982), but uses a different and more efficient smoothing algorithm. Let $\mathbf{x} = [\mathbf{x}_1', \ldots, \mathbf{x}_n]'$, $\mathbf{f} = [\mathbf{f}_0', \mathbf{f}_1', \ldots, \mathbf{f}_n']'$, and let $g(\cdot)$ denote the Gaussian probability density function. The factor model formulated in (2) is such that $\ln g(\mathbf{f}|\mathbf{x};\Xi) = \ln g(\mathbf{x},\mathbf{f};\Xi) - \ln g(\mathbf{x};\Xi)$, where the first term on the right hand side is the joint probability density function of the observations and the factors, also known as the complete data likelihood, and the subtrahend is the likelihood, $\log L(\Xi) = \ln g(\mathbf{x};\Xi)$, of the observed data.

The complete data likelihood can be evaluated as follows: $\ln g(\mathbf{x},\mathbf{f};\Xi) = \ln g(\mathbf{x}|\mathbf{f};\Xi) + \ln g(\mathbf{f};\Xi)$, where $\ln g(\mathbf{x}|\mathbf{f};\Xi) = \sum_{t=1}^n \ln g(\mathbf{x}_t|\mathbf{f}_t)$, and $\ln g(\mathbf{f};\Xi) = \sum_{t=1}^n \ln g(\mathbf{f}_t|\mathbf{f}_{t-1};\Xi) + \ln g(\mathbf{f}_0;\Xi)$. Thus, from (2),

$$\begin{aligned} \ln g(\mathbf{x},\mathbf{f};\Xi) \;=\; & -\tfrac{1}{2}\left[n \ln|\mathbf{\Psi}| + \mathrm{tr}\left\{\mathbf{\Psi}^{-1}\textstyle\sum_{t=1}^n(\mathbf{x}_t - \mathbf{\Theta}\mathbf{f}_t)(\mathbf{x}_t - \mathbf{\Theta}\mathbf{f}_t)'\right\}\right] \\ & -\tfrac{1}{2}\left[n \ln|\mathbf{\Sigma}_\eta| + \mathrm{tr}\left\{\mathbf{\Sigma}_\eta^{-1}\textstyle\sum_{t=1}^n(\mathbf{f}_t - \mathbf{\Phi}\mathbf{f}_{t-1})(\mathbf{f}_t - \mathbf{\Phi}\mathbf{f}_{t-1})'\right\}\right] \\ & -\tfrac{1}{2}\left[\ln|\mathbf{P}_0| + \mathrm{tr}\left\{\mathbf{P}_0^{-1}\mathbf{f}_0\mathbf{f}_0'\right\}\right] \end{aligned}$$

where $\mathbf{P}_0$ satisfies the matrix equation $\mathbf{P}_0 = \mathbf{\Phi}\mathbf{P}_0\mathbf{\Phi}' + \mathbf{\Sigma}_\eta$.

Given an initial parameter value, Ξ^*, the EM algorithm iteratively maximizes, with respect to Ξ, the intermediate quantity (Dempster *et al.*, 1977):

$$Q(\Xi;\Xi^*) = \mathrm{E}_{\Xi^*}\left[\ln g(\mathbf{x},\mathbf{f};\Xi)\right] = \int \ln g(\mathbf{x},\mathbf{f};\Xi) g(\mathbf{f}|\mathbf{x};\Xi^*) d\mathbf{f},$$

which is interpreted as the expectation of the complete data log-likelihood with respect to $g(\mathbf{f}|\mathbf{x};\Xi^*)$, which is the conditional probability density function of the unobservable states, given the observations, evaluated using $\mathbf{\Xi}^*$. Now,

$$\begin{aligned} Q(\mathbf{\Xi};\mathbf{\Xi}^*) \;\; = \;\; & -\tfrac{1}{2}\left[n\ln|\mathbf{\Psi}| + \text{tr}\left\{\mathbf{\Psi}^{-1}\textstyle\sum_{t=1}^{n}\left[(\mathbf{x}_t - \mathbf{\Theta}\tilde{\mathbf{f}}_{t|n})(\mathbf{x}_t - \mathbf{\Theta}\tilde{\mathbf{f}}_{t|n})' + \mathbf{\Theta}\mathbf{P}_{t|n}\mathbf{\Theta}'\right]\right\}\right] \\ & -\tfrac{1}{2}\left[n\ln|\mathbf{\Sigma}_\eta| + \text{tr}\left\{\mathbf{\Sigma}_\eta^{-1}(\mathcal{S}_f - \mathcal{S}_{f,f-1}\mathbf{\Theta}' - \mathbf{\Theta}\mathcal{S}'_{f,f-1} + \mathbf{\Theta}\mathcal{S}_{f-1}\mathbf{\Theta}')\right\}\right] \\ & -\tfrac{1}{2}\left[\ln|\mathbf{P}_0| + \text{tr}\left\{\mathbf{P}_0^{-1}(\tilde{\mathbf{f}}_{0|n}\tilde{\mathbf{f}}'_{0|n} + \mathbf{P}_{0|n})\right\}\right] \end{aligned}$$

where $\tilde{\mathbf{f}}_{t|n} = \text{E}(\mathbf{f}_t|\mathbf{x};\mathbf{\Xi}^{(j)})$, $\mathbf{P}_{t|n} = \text{Var}(\mathbf{f}_t|\mathbf{x};\mathbf{\Xi}^{(j)})$, and

$$\mathcal{S}_f = \left[\sum_{t=1}^{n}\left(\mathbf{P}_{t|n} + \tilde{\mathbf{f}}_{t|n}\tilde{\mathbf{f}}'_{t|n}\right)\right],$$

$$\mathcal{S}_{f-1} = \left[\sum_{t=1}^{n}\left(\mathbf{P}_{t-1|n} + \tilde{\mathbf{f}}_{t-1|n}\tilde{\mathbf{f}}'_{t-1|n}\right)\right], \mathcal{S}_{f,f-1} = \left[\sum_{t=1}^{n}\left(\mathbf{P}_{t,t-1|n} + \tilde{\mathbf{f}}_{t|n}\tilde{\mathbf{f}}'_{t-1|n}\right)\right].$$

These quantities are evaluated with the support of the Kalman filter and smoother (KFS, see below), adapted to the state space model (2) with parameter values $\mathbf{\Xi}^*$. Also, $\mathbf{P}_{t,t-1|n} = \text{Cov}(\mathbf{f}_t, \mathbf{f}_{t-1}|\mathbf{x};\mathbf{\Xi}^*)$ is computed using the output of the KFS recursions, as it will be detailed below.

Dempster *et al.* (1977) show that the parameter estimates maximising the log-likelihood $\log L(\mathbf{\Xi})$, can be obtained by a sequence of iterations, each consisting of an expectation step (E-step) and a maximization step (M-step), that aim at locating a stationary point of $Q(\mathbf{\Xi};\mathbf{\Xi}^*)$. At iteration j, given the estimate $\mathbf{\Xi}^{(j)}$, the E-step deals with the evaluation of $Q(\mathbf{\Xi};\mathbf{\Xi}^{(j)})$; this is carried out with the support of the KFS applied to the state space representation (2) with hyperparameters $\mathbf{\Xi}^{(j)}$.

The M-step amounts to choosing a new value $\mathbf{\Xi}^{(j+1)}$, so as to maximize with respect to $\mathbf{\Xi}$ the criterion $Q(\mathbf{\Xi};\mathbf{\Xi}^{(j)})$, i.e., $Q(\mathbf{\Xi}^{(j+1)};\mathbf{\Xi}^{(j)}) \geq Q(\mathbf{\Xi}^{(j)};\mathbf{\Xi}^{(j)})$. The maximization is in closed form, if we assume that $\mathbf{P}_0$ is an independent unrestricted parameter. Actually, the latter depends on the matrices $\mathbf{\Phi}$ and $\mathbf{\Sigma}_\eta$, but we will ignore this fact, as it is usually done. For the loadings matrix the M-step consists of maximizing $Q(\mathbf{\Xi};\mathbf{\Xi}^{(j)})$ with respect to $\mathbf{\Theta}$, subject to subject to the identification constraints: $\mathbf{C}'\mathbf{\Theta} = \mathbf{I}_K$, where $\mathbf{C}' = [\mathbf{I}_K, \mathbf{0}]$. Denoting the unconstrained estimate by

$$\hat{\mathbf{\Theta}}_U^{(j+1)} = \left(\sum_{t=1}^{n}\mathbf{x}_t\tilde{\mathbf{f}}'_{t|n}\right)\mathcal{S}_f^{-1},$$

the constrained estimate is (Magnus and Neudecker, 2007)

$$\hat{\mathbf{\Theta}}^{(j+1)} = \left[(\mathbf{I}_n - \mathbf{C}\mathbf{C}')\sum_{t=1}^{n}\mathbf{x}_t\tilde{\mathbf{f}}'_{t|n} + \mathbf{C}\mathcal{S}_f\right]\mathcal{S}_f^{-1} = (\mathbf{I}_n - \mathbf{C}\mathbf{C}')\hat{\mathbf{\Theta}}_U^{(j+1)} + \mathbf{C} = \hat{\mathbf{\Theta}}_U^{(j+1)} - \mathbf{C}(\mathbf{C}'\hat{\mathbf{\Theta}}_U^{(j+1)} - \mathbf{I}_K)$$

since $\mathbf{C}'\mathbf{C} = \mathbf{I}_K$.

The $(j+1)$ update of the matrix $\mathbf{\Psi}$ is given by

$$\hat{\mathbf{\Psi}}^{(j+1)} = \text{diag}\left\{\frac{1}{n}\sum_{t=1}^{n}\left[\mathbf{x}_t\mathbf{x}'_t - \hat{\mathbf{\Theta}}^{(j+1)}\tilde{\mathbf{f}}_{t|n}\mathbf{x}'_t\right]\right\}.$$

Further, we have:

$$\hat{\mathbf{\Phi}}^{(j+1)} = \mathcal{S}_{f,f-1}\mathcal{S}_{f-1}^{-1}, \quad \hat{\mathbf{\Sigma}}_\eta^{(j+1)} = \frac{1}{n}\left(\mathcal{S}_f - \hat{\mathbf{\Phi}}^{(j+1)}\mathcal{S}'_{f,f-1}\right).$$

In the above expressions $\tilde{\mathbf{f}}_{t|n} = \mathrm{E}(\mathbf{f}_t|\mathbf{x};\boldsymbol{\Xi}^{(j)})$ and $\mathbf{P}_{t|n} = \mathrm{Var}(\mathbf{f}_t|\mathbf{x};\boldsymbol{\Xi}^{(j)})$ are computed by the KFS. Also, $\mathbf{P}_{t,t-1|n} = \mathrm{Cov}(\mathbf{f}_t, \mathbf{f}_{t-1}|\mathbf{x};\boldsymbol{\Xi}^{(j)})$ is computed from the KFS recursions, as we now explain. Defining the initial values $\tilde{\mathbf{f}}_{1|0} = \mathbf{0}$, and $\mathbf{P}_{1|0} = \mathbf{P}_0$, the Kalman filter is given by the following recursive formulae and definitions, for $t = 1, \ldots, n$:

$$\begin{array}{lll} \mathbf{v}_t = \mathbf{x}_t - \boldsymbol{\Theta}\tilde{\mathbf{f}}_{t|t-1}, & \mathbf{F}_t = \boldsymbol{\Theta}\mathbf{P}_{t|t-1}\boldsymbol{\Theta}' + \boldsymbol{\Psi}, & \mathbf{K}_t = \boldsymbol{\Phi}\mathbf{P}_{t|t-1}\boldsymbol{\Theta}'\mathbf{F}_t^{-1}, \\ \tilde{\mathbf{f}}_{t+1|t} = \boldsymbol{\Phi}\tilde{\mathbf{f}}_{t|t-1} + \mathbf{K}_t\mathbf{v}_t, & \mathbf{P}_{t+1|t} = \boldsymbol{\Phi}\mathbf{P}_{t|t-1}\boldsymbol{\Phi}' + \boldsymbol{\Sigma}_\eta - \mathbf{K}_t\mathbf{F}_t\mathbf{K}_t' & \end{array} \tag{3}$$

here, $\mathbf{v}_t = \mathbf{x}_t - \mathrm{E}(\mathbf{x}_t|\mathbf{x}^{t-1};\boldsymbol{\Xi})$, $\mathbf{F}_t = \mathrm{Var}(\mathrm{v}_t|\mathbf{x}^{t-1};\boldsymbol{\Xi})$, $\tilde{\mathbf{f}}_{t|t-1} = \mathrm{E}(\mathbf{f}_t|\mathbf{x}^{t-1};\boldsymbol{\Xi})$, $\mathbf{P}_{t|t-1} = \mathrm{Var}(\mathbf{f}_t|\mathbf{x}^{t-1};\boldsymbol{\Xi})$.

The smoothed estimates $\tilde{\mathbf{f}}_{t|n} = \mathrm{E}(\boldsymbol{\alpha}_t|\mathbf{x};\boldsymbol{\Xi})$, and their covariance matrix $\mathbf{P}_{t|n} = \mathrm{E}[(\mathbf{f}_t - \tilde{\mathbf{f}}_{t|n})(\mathbf{f}_t - \tilde{\mathbf{f}}_{t|n})'|\mathbf{x};\boldsymbol{\Xi}]$ are computed by the following backwards recursive formulae, given by Bryson and Ho (1969) and de Jong (1989), starting at $t = n$, with initial values $\mathbf{r}_n = 0, \mathbf{R}_n = \mathbf{0}$ and $\mathbf{N}_n = 0$: for $t = n-1, \ldots, 1$,

$$\begin{array}{rcll} \mathbf{r}_{t-1} & = & \mathbf{L}_t'\mathbf{r}_t + \mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{v}_t, & \mathbf{M}_{t-1} = \mathbf{L}_t'\mathbf{M}_t\mathbf{L}_t + \mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{Z}_t, \\ \tilde{\mathbf{f}}_{t|n} & = & \tilde{\mathbf{f}}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{r}_{t-1}, & \mathbf{P}_{t|n} = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{M}_{t-1}\mathbf{P}_{t|t-1}. \end{array} \tag{4}$$

where $\mathbf{L}_t = \boldsymbol{\Phi} - \mathbf{K}_t\boldsymbol{\Theta}'$.

The smoothed estimates of the disturbances are given by $\tilde{\boldsymbol{\eta}}_{t|n} = \mathrm{E}(\boldsymbol{\eta}_t|\mathbf{x};\boldsymbol{\Xi}) = \boldsymbol{\Sigma}_\eta\mathbf{r}_{t-1}$, and $\tilde{\boldsymbol{\xi}}_{t|n} = \mathrm{E}(\boldsymbol{\xi}_t|\mathbf{x};\boldsymbol{\Xi}) = \boldsymbol{\Psi}\left[\mathbf{F}_t^{-1}\mathbf{v}_t + \mathbf{K}_t'\mathbf{r}_t\right]$. Indeed, the vector $\mathbf{r}_{t-1}$ computes $\boldsymbol{\Sigma}_\eta^{-1}\mathrm{Cov}(\boldsymbol{\eta}_t, \mathbf{x})\mathrm{Var}(\mathbf{x})^{-1}(\mathbf{x}-\mathrm{E}(\mathbf{x})) = \boldsymbol{\Sigma}_\eta^{-1}\sum_{j=t}^{n}\mathrm{Cov}(\boldsymbol{\eta}_t, \mathbf{v}_j)\mathbf{F}_j^{-1}\mathbf{v}_j$. The matrix $\mathbf{M}_{t-1}$ computes $\mathrm{Var}(\mathbf{r}_{t-1}) = \boldsymbol{\Sigma}_\eta^{-1}\mathrm{Cov}(\boldsymbol{\eta}_t, \mathbf{x})\mathrm{Var}(\mathbf{x})^{-1}\mathrm{Cov}(\mathbf{x}, \boldsymbol{\eta}_t)\boldsymbol{\Sigma}_\eta^{-1}$. The derivation of these expressions follows Koopman (1993).

Finally,

$$\mathbf{P}_{t,t-1|n} = \mathrm{Cov}(\mathbf{f}_t, \mathbf{f}_{t-1}|\mathbf{x}) = \boldsymbol{\Phi}\mathbf{P}_{t-1|n} - \boldsymbol{\Sigma}_\eta\mathbf{M}_{t-1}\mathbf{L}_{t-1}\mathbf{P}_{t-1|t-2}.$$

The proof of this result is given below:

$$\begin{array}{rcl} \mathrm{Cov}(\mathbf{f}_t, \mathbf{f}_{t-1}|\mathbf{x}) & = & \boldsymbol{\Phi}\mathrm{Var}(\mathbf{f}_{t-1}|\mathbf{x}) + \mathrm{Cov}(\boldsymbol{\eta}_t, \mathbf{f}_{t-1}|\mathbf{x}) \\ & = & \boldsymbol{\Phi}\mathbf{P}_{t-1|n} + \mathrm{Cov}(\boldsymbol{\eta}_t, \mathbf{f}_{t-1}) - \mathrm{Cov}(\tilde{\boldsymbol{\eta}}_{t|n}, \tilde{\mathbf{f}}_{t-1|n}) \\ & = & \boldsymbol{\Phi}\mathbf{P}_{t-1|n} - \mathrm{Cov}(\tilde{\boldsymbol{\eta}}_{t|n}, \tilde{\mathbf{f}}_{t-1|n}) \\ & = & \boldsymbol{\Phi}\mathbf{P}_{t-1|n} - \mathrm{Cov}(\boldsymbol{\Sigma}_\eta\mathbf{r}_{t-1}, \tilde{\mathbf{f}}_{t-1|t-2} + \mathbf{P}_{t-1|t-2}\mathbf{r}_{t-2}) \\ & = & \boldsymbol{\Phi}\mathbf{P}_{t-1|n} - \mathrm{Cov}(\boldsymbol{\Sigma}_\eta\mathbf{r}_{t-1}, \mathbf{P}_{t-1|t-2}\mathbf{L}_{t-1}'\mathbf{r}_{t-1}) \\ & = & \boldsymbol{\Phi}\mathbf{P}_{t-1|n} - \boldsymbol{\Sigma}_\eta\mathbf{M}_{t-1}\mathbf{L}_{t-1}\mathbf{P}_{t-1|t-2}. \end{array}$$

The covariances for smoothed estimates were derived by de Jong and Mackinnon (1988). Our derivation is different since it is based on the output of the Bryson and Ho (1969) and de Jong (1989) smoothing algorithm, which is more efficient with respect to that considered by Shumway and Stoffer (1982) and de Jong and Mackinnon (1988).

4.2 Principal components analysis

The static principal component estimator minimizes, with respect to $\hat{\mathbf{f}}_t, t = 1, \ldots, n$, and $\hat{\boldsymbol{\Theta}}$, the nonlinear least squares criterion (see Stock and Watson, 2002b, and Bai, 2003):

$$\sum_t (\mathbf{x}_t - \hat{\boldsymbol{\Theta}}\hat{\mathbf{f}}_t)'(\mathbf{x}_t - \hat{\boldsymbol{\Theta}}\hat{\mathbf{f}}_t),$$

subject to the normalisations $\hat{\boldsymbol{\Theta}}'\hat{\boldsymbol{\Theta}} = \mathbf{I}_K$ and $n^{-1}\sum_t \hat{\mathbf{f}}_t\hat{\mathbf{f}}_t' = \mathrm{diag}\{\lambda_k > 0, k = 1, \ldots, K\}$, which altogether define the K^2 restrictions that are required for exact identification. The solution yields $\hat{\boldsymbol{\Theta}}$ as the matrix whose columns are formed from the first K eigenvectors of the covariance matrix $n^{-1}\sum_t \mathbf{x}_t\mathbf{x}_t'$, corresponding to the K largest eigenvalues, and $\hat{\mathbf{f}}_t = \hat{\boldsymbol{\Theta}}'\mathbf{x}_t$. Notice that this approach treats the factors as fixed parameters, and thus their estimator is coincident with the first K principal components. However, for $n, N \to \infty$ this is asymptotically equivalent to the Wiener-Kolmogorov estimator of the factors, in that the estimation mean square error converges to zero. This is formally shown under different assumptions in Stock and Watson (2002b), Bai (2003) and Doz, Giannone and Reichlin (2007, section 3).

The principal components $\hat{\mathbf{f}}_t$, can be used for estimating the VAR coefficients and disturbance covariance matrix:

$$\hat{\boldsymbol{\Phi}} = \sum_{t=2}^{n} \hat{\mathbf{f}}_t\hat{\mathbf{f}}_{t-1}' \left(\sum_{t=2}^{n} \hat{\mathbf{f}}_{t-1}\hat{\mathbf{f}}_{t-1}'\right)^{-1}, \quad \hat{\boldsymbol{\Sigma}}_\eta = \frac{1}{n-1}\left(\sum_{t=2}^{n}\hat{\mathbf{f}}_t\hat{\mathbf{f}}_t' - \hat{\boldsymbol{\Phi}}\sum_{t=2}^{n}\hat{\mathbf{f}}_{t-1}\hat{\mathbf{f}}_t'\right).$$

Finally, $\hat{\boldsymbol{\Psi}} = \text{diag}\left\{\frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t\mathbf{x}_t' - \hat{\boldsymbol{\Theta}}\hat{\mathbf{f}}_t\mathbf{x}_t'\right\}$. The consistency of the estimator of $\boldsymbol{\Xi}$ based on PCA has been shown by Bai (2003) and Forni et al. (2005) for $n, N \to \infty$ under different settings. Giannone, Reichlin and Sala (2005) use a two step estimator of the factors, such that the parameters $\boldsymbol{\Xi}$ are estimated by PCA and the factors by the KFS. Doz, Giannone and Reichlin (2007) prove the consistency of such estimator.

For comparison with the EM estimates, the PCA solution will be rotated. In particular, if $\hat{\boldsymbol{\Theta}}_{(K)}$ denotes the first row block of $\hat{\boldsymbol{\Theta}}$, so that $\hat{\boldsymbol{\Theta}} = [\hat{\boldsymbol{\Theta}}_{(K)}', \hat{\boldsymbol{\Theta}}_{(U)}']'$, we shall consider the estimate

$$\tilde{\boldsymbol{\Theta}} = \hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Theta}}_{(K)}^{-1} = \begin{bmatrix} \mathbf{I}_K \\ \hat{\boldsymbol{\Theta}}_{(U)}\hat{\boldsymbol{\Theta}}_{(K)}^{-1} \end{bmatrix},$$

which enforces the restriction that the upper block is the identity matrix. Consequently, the estimates of the VAR coefficient matrix and disturbance variance matrix are, respectively, $\tilde{\boldsymbol{\Phi}} = \hat{\boldsymbol{\Theta}}_{(K)}\hat{\boldsymbol{\Phi}}\hat{\boldsymbol{\Theta}}_{(K)}^{-1}$ and $\tilde{\boldsymbol{\Sigma}}_\eta = \hat{\boldsymbol{\Theta}}_{(K)}\hat{\boldsymbol{\Sigma}}_\eta\hat{\boldsymbol{\Theta}}_{(K)}'$.

Another possibility is to base estimation on weighted principal components, where the weights are proportional to the inverse of the standard deviation of the idiosyncratic component; this is discussed in Boivin and Ng (2004) and Forni et al. (2005), but will not be explored no further.

5 Temporal aggregation

The N time series y_{it} are available at different frequencies of observation. In particular, the first block of $N_1 = 17$ time series, GDP and its main components, are quarterly. Since $Y_{it}, 1, \ldots, N_1$, is subject to temporal aggregation, we observe the quarterly totals:

$$Y_{i\tau} = \sum_{i=1}^{3} Y_{i,3\tau-i}, \quad \tau = 1, 2, \ldots, [(n+1)/3], \tag{5}$$

where $[\cdot]$ is the integer part of the argument.

For the statistical treatment it is useful to convert temporal aggregation into a systematic sampling problem; this can be done by constructing a cumulator variable, generated as a time-varying first order autoregression (see Harvey, 1989, and Harvey and Chung, 2000):

$$\begin{aligned} Y_{it}^c &= \rho_t Y_{i,t-1}^c + Y_{it}, \quad t = 0, \ldots, n \\ &= \rho_t Y_{i,t-1}^c + h_i(y_{it}) \end{aligned} \tag{6}$$

where $h_i(\cdot)$ is the Box-Cox inverse transformation,

$$h_i(y_{it}) = \begin{cases} (1+\lambda_i y_{it})^{1/\lambda_i}, & \lambda_i \neq 0, \\ \exp(y_{it}), & \lambda_i = 0, \end{cases}$$

and ρ_t is the cumulator coefficient, equal to zero for t corresponding to the first month in the quarter and 1 otherwise:

$$\rho_t = \begin{cases} 0 & t = 3(\tau-1), \quad \tau = 1, \dots, [(n+1)/3] \\ 1 & \text{otherwise} . \end{cases}$$

The cumulator (6) is nothing more than a recursive implementation of the temporal aggregation rule (5). Only a systematic sample of the cumulator variable Y_{it}^c is available; in particular, if the sample period starts with the first month of the quarter at $t = 0$, the observed end of quarter values occur at times $t = 3\tau - 1, \tau = 1, 2, \dots, [(n+1)/3]$

In the case of the logarithmic transformation ($\lambda_i = 0$), $Y_{i0}^c = \exp y_{i0}$, $Y_{i1}^c = \exp(y_{i0}) + \exp(y_{i1})$, $Y_{i2}^c = \exp(y_{i0}) + \exp(y_{i1}) + \exp(y_{i2})$, $Y_{i3}^c = \exp(y_{i3})$, $Y_{i4}^c = \exp(y_{i3}) + \exp(y_{i4})$, $Y_{i5}^c = \exp(y_{i3}) + \exp(y_{i4}) + \exp(y_{i5})$, ...Only the values $Y_{i2}^c, Y_{i5}^c, \dots$ are observed, while the intermediate ones will be missing. It it important to remark that in general, when the Box-Cox transformation parameter is different from one, the quarterly totals are a nonlinear function of the underlying (unobserved) monthly values y_{it} (e.g. the sum of the exponentials of three consecutive values). Now, since we postulate that the first differences Δy_{it} are stationary and they have a linear factor model representation, the temporal aggregation constraints are nonlinear. In other words, we observe $Y_{i\tau}^c = Y_{i,3\tau-1} + Y_{i,3\tau-2} + Y_{i,3\tau-3}$, but the linear model is formulated in terms of the unobserved $y_{i,3\tau-i}, i = 1, 2, 3$, which are the Box-Cox power transformation of $Y_{i,3\tau-i}$. Hence, temporal aggregation yields a nonlinear observational constraint.

6 Nonlinear Smoothing

Conditional on Ξ, we face the problem of estimating the factors $\mathbf{f}_t$ and the missing values $\mathbf{y}_{it}, i = 1, \dots, N_1$, from the available information, which consists of $Y_{it}^c, i = 1, \dots, N_1, t = 3\tau - 1, \tau = 1, 2, \dots, [(n+1)/3]$, for the quarterly time series and y_{it} for $i = N_1 + 1, \dots, N$. This is a nonlinear smoothing problem that can be solved by iterating the Kalman filter and smoother adapted to a sequentially linearized state space model.

The estimation is carried out by an iterative algorithm which is a *sequential linear constrained* method for solving a constrained nonlinear optimization problem; see Gill et al. (1989), section 7. This method has been applied to nonlinear aggregation in mixed models Proietti (2006) and to temporal disaggregation by Proietti and Moauro (2006).

Let us partition the vectors $\mathbf{Y}_t = [\mathbf{Y}_{1t}', \mathbf{Y}_{2t}']'$, $\mathbf{y}_t = [\mathbf{y}_{1t}', \mathbf{y}_{2t}']'$, such that $\mathbf{Y}_t = \mathbf{h}(\mathbf{y}_t)$ is the inverse Box-Cox transform of $\mathbf{y}_t$, $\Delta\mathbf{y}_t = [\Delta\mathbf{y}_{1t}', \Delta\mathbf{y}_{2t}']'$, $\mathbf{x}_t = [\mathbf{x}_{1t}', \mathbf{x}_{2t}']'$, $\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2]'$, and the matrices $\mathbf{D} = \text{diag}(\mathbf{D}_1, \mathbf{D}_2)$, $\boldsymbol{\Theta} = [\boldsymbol{\Theta}_1', \boldsymbol{\Theta}_2']'$, $\boldsymbol{\Psi} = \text{diag}(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2)$, where the subscript 1 indexes the national accounts series, and the dimension of the blocks are respectively N_1 and N_2. Further, define $\boldsymbol{\xi} = [\boldsymbol{\xi}_1', \dots, \boldsymbol{\xi}_n']'$, i.e. the stack of the idiosyncratic disturbances.

If $\mathbf{x}_t$ were fully observed and Ξ were known, the KFS would yield the values of $\mathbf{f}$ and $\boldsymbol{\xi}$ that maximise the complete data likelihood $g(\mathbf{x}, \mathbf{f}; \Xi) = g(\mathbf{x}|\mathbf{f}; \Xi)g(\mathbf{f}; \Xi)$. Now, $\mathbf{x}_{1t}, t = 1, \dots, n$, is not available, but we observe a systematic sample of the cumulator

$$\begin{aligned} \mathbf{Y}_{1t}^c &= \rho_t \mathbf{Y}_{1,t-1}^c + \mathbf{Y}_{1t}, \\ &= \rho_t \mathbf{Y}_{1,t-1}^c + \mathbf{h}(\mathbf{y}_{1t}), \end{aligned}$$

and $\mathbf{x}_{1t}$ is related to $\mathbf{y}_{1t}$ by $\mathbf{x}_{1t} = \mathbf{D}_1^{-1}(\Delta\mathbf{y}_{1t} - \boldsymbol{\mu}_1)$.

The smoothing problem is now to obtain the values $\mathbf{f}$ and $\boldsymbol{\xi}$ that maximise the complete data likelihood $g(\mathbf{x}, \mathbf{f}; \boldsymbol{\Xi})$, subject to the nonlinear observational constraints that we observe a systematic sample of $\mathbf{Y}_{1t}^c = \rho_t \mathbf{Y}_{1,t-1}^c + \mathbf{h}(\mathbf{y}_{1t})$, and $\mathbf{x}_{1t} = \mathbf{D}_1^{-1}(\Delta\mathbf{y}_{1t} - \boldsymbol{\mu}_1)$.

The optimisation problem is handled with the support of the KFS. Each time the observation constraint is linearised around a trial value by a first order Taylor series expansion; this operation yields a linear state space model and the corresponding KFS provides a new trial value for the disaggregate series. This sequence of linearisations is iterated until convergence and the end result is a set of disaggregate monthly estimates $\mathbf{Y}_1$ and factor scores which incorporate the temporal aggregation constraints. As a by-product, disaggregate (monthly) estimates of the missing values $\mathbf{x}_{1t}$ and thus of $\mathbf{y}_{1t}$ and $\mathbf{Y}_{it}$ will be made available.

The linearisation operates as follows. Let $\mathbf{y}_{1t}^*$ denote a trial estimate of the Box-Cox transformed disaggregate series, and $\mathbf{Y}_{1t}^* = \mathbf{h}(\mathbf{y}_{1t}^*)$. Linearising the cumulator around it, using the first order Taylor approximation, yields

$$\mathbf{Y}_{1t}^c = \rho_t \mathbf{Y}_{1,t-1}^c + \mathbf{h}(\mathbf{y}_{it}^*) + \mathbf{U}_{1t}^*(\mathbf{y}_{1t} - \mathbf{y}_{1t}^*)$$

where the $N_1 \times N_1$ matrix $\mathbf{U}_{1t}^*$ is a diagonal matrix with the derivatives of the inverse Box-Cox transformation on the main diagonal

$$\mathbf{U}_{1t}^* = \text{diag}\left(\left. \frac{dh_i(y_{it})}{dy_{it}} \right|_{y_{it}=y_{it}^*}, i = 1, 2, \dots, N_1 \right)$$

in the case $\lambda_i = 0, i = 1, \dots, N_1$, (logarithmic transformation for all the variables), $h_i(y_{it}) = \exp(y_{it})$ and $\mathbf{U}_{1t}^* = \text{diag}(\exp(\mathbf{y}_{1t}^*))$.

When $\mathbf{y}_{1t}$ is difference stationary, as in our case, writing $\mathbf{y}_{1t} = \mathbf{y}_{1,t-1} + \Delta\mathbf{y}_{1t} = \mathbf{y}_{1,t-2} + \Delta\mathbf{y}_{1,t-1} + \Delta\mathbf{y}_{1t}$, replacing

$$\begin{aligned} \Delta\mathbf{y}_{1t} &= \boldsymbol{\mu}_1 + \mathbf{D}_1\mathbf{x}_{1t} \\ &= \boldsymbol{\mu}_1 + \mathbf{D}_1(\boldsymbol{\Theta}_1\mathbf{f}_t + \boldsymbol{\xi}_{1t}) \\ &= \boldsymbol{\mu}_1 + \mathbf{D}_1(\boldsymbol{\Theta}_1\boldsymbol{\Phi}\mathbf{f}_{t-1} + \boldsymbol{\Theta}_1\boldsymbol{\eta}_t + \boldsymbol{\xi}_{1t}), \end{aligned}$$

and rearranging, enables us to express $\mathbf{Y}_{1t}^c$ as a time-varying linear combination of $\mathbf{Y}_{1,t-1}^c, \mathbf{y}_{1,t-2}, \Delta\mathbf{y}_{1,t-1}, \mathbf{f}_{t-1}$, which will constitute the elements of the state vector at time $t-1$, denoted $\boldsymbol{\alpha}_{t-1}$:

$$\begin{aligned} \mathbf{Y}_{1t}^c = \; & \rho_t \mathbf{Y}_{1,t-1}^c + \mathbf{U}_{1t}^*(\mathbf{y}_{1,t-2} + \Delta\mathbf{y}_{1,t-1} + \mathbf{D}_1\boldsymbol{\Theta}_1\boldsymbol{\Phi}\mathbf{f}_{t-1}) + \mathbf{Y}_{1t}^* - \mathbf{U}_{1t}^*\mathbf{y}_{1t}^* + \\ & \mathbf{U}_{1t}^*\boldsymbol{\mu}_1 + \mathbf{U}_{1t}^*\mathbf{D}_1(\boldsymbol{\Theta}_1\boldsymbol{\eta}_t + \boldsymbol{\xi}_{1t}). \end{aligned}$$

When $h_i(\cdot) = \exp(\cdot), \forall i$, (i.e. in the case $\lambda_i = 0, i = 1, \dots, N_1$), $\mathbf{y}_{1t}^* = \log(\mathbf{Y}_{1t}^*)$, $\mathbf{U}_{1t}^* = \text{diag}\,(\exp(\mathbf{y}_{1t}^*)) = \text{diag}\,(\mathbf{Y}_{1t}^*))$, and $\mathbf{h}(\mathbf{y}_{1t}^*) - \mathbf{U}_{1t}^*\mathbf{y}_{1t}^* = \mathbf{Y}_{1t}^{c*} - \mathbf{U}_{1t}^* \log(\mathbf{Y}_{1t}^*)$.

6.1 State space representation

The state space representation is conveniently formulated for the vector $\mathbf{y}_t^\dagger$, given by

$$\mathbf{y}_t^\dagger = \begin{bmatrix} \mathbf{Y}_{1t}^c \\ \Delta\mathbf{y}_{2t} \end{bmatrix}, t = 1, 2, \dots, n,$$

whereas for $t = 0$, $\mathbf{y}_0^\dagger = \mathbf{Y}_{10}^c$. The length of the observation vector varies with time and will be denoted by N_t.

The measurement equation is

$$\mathbf{y}_t^\dagger = \mathbf{Z}_t\boldsymbol{\alpha}_t + \mathbf{c}_t + \mathbf{G}_t\boldsymbol{\xi}_{2t}, \boldsymbol{\xi}_{2t} \sim \text{NID}(\mathbf{0}, \boldsymbol{\Psi}_2), \tag{7}$$

where $\boldsymbol{\xi}_{2t}$ is the vector of idiosyncratic disturbances of the factor model for the second block of time series, which contains those time series that are fully observed at the monthly frequency. At time $t = 0$ the measurement equation is formulated in terms of the N_1 elements $\mathbf{Y}_{1,0}^c$:

$$\mathbf{y}_0^\dagger = \mathbf{Y}_{1,0}^c, \mathbf{Z}_0 = \left[\ \mathbf{I}_{N_1}, \mathbf{0}, \mathbf{0}, \mathbf{0}\ \right], \mathbf{c}_0 = \mathbf{0}, \mathbf{G}_0 = \mathbf{0}.$$

For all times times $t \geq 1$,

$$\mathbf{y}_t^\dagger = \begin{bmatrix} \mathbf{Y}_{1,t}^c \\ \Delta\mathbf{y}_{2t} \end{bmatrix}, \mathbf{Z}_t = \begin{bmatrix} \mathbf{I}_{N1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{D}_2\boldsymbol{\Theta}_2 \end{bmatrix}, \mathbf{c}_t = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\mu}_2 \end{bmatrix}, \mathbf{G}_t = \begin{bmatrix} \mathbf{0} \\ \mathbf{D}_2 \end{bmatrix},$$

It should be recalled that only a systematic sample of $\mathbf{Y}_{1t}^c$ is available at times $3\tau - 1, \tau = 1, \ldots, [(n+1)/3]$, and thus the measurement equation is subject to missing values.

The transition equation is defined as

$$\boldsymbol{\alpha}_t = \mathbf{T}_t\boldsymbol{\alpha}_{t-1} + \mathbf{d}_t + \mathbf{H}_t\boldsymbol{\omega}_t, \quad t = 1, \ldots, n,$$

where the state and the disturbance vectors are

$$\boldsymbol{\alpha}_t = \begin{bmatrix} \mathbf{Y}_{1,t}^c \\ \mathbf{y}_{1,t-1} \\ \Delta\mathbf{y}_{1t} \\ \mathbf{f}_t \end{bmatrix}, \boldsymbol{\omega}_t = \begin{bmatrix} \boldsymbol{\xi}_{1t} + \boldsymbol{\Theta}_1\boldsymbol{\eta}_t \\ \boldsymbol{\eta}_t \end{bmatrix},$$

and

$$\mathbf{T}_t = \begin{bmatrix} \rho_t\mathbf{I}_{N1} & \mathbf{U}_{1t}^* & \mathbf{U}_{1t}^* & \mathbf{U}_{1t}^*\mathbf{D}_1\boldsymbol{\Theta}_1\boldsymbol{\Phi} \\ \mathbf{0} & \mathbf{I}_N & \mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{D}_1\boldsymbol{\Theta}_1\boldsymbol{\Phi} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\Phi} \end{bmatrix}, \mathbf{H}_t = \begin{bmatrix} \mathbf{U}_{1t}^*\mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_K \end{bmatrix}, \mathbf{d}_t = \begin{bmatrix} \mathbf{Y}_{1t}^* - \mathbf{Z}_{it}^*\mathbf{y}_{1t}^* + \mathbf{U}_{1t}^*\boldsymbol{\mu}_1 \\ \mathbf{0} \\ \boldsymbol{\mu}_1 \\ \mathbf{0} \end{bmatrix}.$$

It must be remarked that $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{D}_1, \mathbf{D}_2$, and the matrices $\boldsymbol{\Theta}, \boldsymbol{\Psi}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}_\eta$ are treated as known quantities.

6.2 Initial conditions

The specification of the state space model is completed by the distribution of the initial state vector $\boldsymbol{\alpha}_0 = \left[\mathbf{Y}_{1,0}^{c\prime}, \mathbf{y}_{1,-1}', \Delta\mathbf{y}_{1,0}', \mathbf{f}_0'\right]'$. The first block is rewritten $\mathbf{Y}_{1,0}^c = f(\mathbf{y}_{1,0})$, as $\rho_0 = 0$; its first order Taylor approximation around the trial value $\mathbf{y}_{1,0}^*$ is

$$\mathbf{Y}_{1,0}^c = f(\mathbf{y}_{1,0}^*) + \mathbf{Z}_{1,0}^*\mathbf{y}_{1,0} - \mathbf{Z}_{1,0}^*\mathbf{y}_{1,0}^*.$$

The first two blocks of the state vector are nonstationary and are initialised by the a vector $\boldsymbol{\beta} = \mathbf{y}_{1,0}$, whereas the last two blocks have a stationary distribution, which depends on $\mathbf{f}_0 \sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma}_f)$, where $\boldsymbol{\Sigma}_f$ solves the matrix equation $\boldsymbol{\Sigma}_f = \boldsymbol{\Phi}\boldsymbol{\Sigma}_f\boldsymbol{\Phi}' + \boldsymbol{\Sigma}_\eta$.

The initial state vector is thus written as:

$$\boldsymbol{\alpha}_0 = \mathbf{A}_{0,0}\boldsymbol{\beta} + \mathbf{a}_{0,0} + \mathbf{H}_0\boldsymbol{\omega}_0, \boldsymbol{\omega}_0 = \begin{bmatrix} \boldsymbol{\xi}_{1,0} + \boldsymbol{\Theta}_1\mathbf{f}_0 \\ \mathbf{f}_0 \end{bmatrix},$$

where

$$\mathbf{a}_{0,0} = \mathbf{d}_0 = \begin{bmatrix} \mathbf{Y}_{1,0}^{c*} - \mathbf{U}_{1,0}^*\mathbf{y}_{1,0}^* + \mathbf{U}_{1,0}^*\boldsymbol{\mu}_1 \\ \mathbf{0} \\ \boldsymbol{\mu}_1 \\ \mathbf{0} \end{bmatrix}, \mathbf{A}_{0,0} = \begin{bmatrix} \mathbf{U}_{1,0}^* \\ \mathbf{I}_{N1} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \mathbf{P}_{0,0} = \mathbf{H}_0\mathrm{Cov}(\boldsymbol{\omega}_0)\mathbf{H}_0'.$$

As far as the vector $\boldsymbol{\beta}$ is concerned, two assumptions can be made: (i) $\boldsymbol{\beta}$ is a fixed unknown vector ($\boldsymbol{\Sigma}_\beta \to 0$); this is suitable if it is deemed that the transition process governing the states has started at time $t = 1$; (ii) $\boldsymbol{\beta}$ is a diffuse random vector, i.e. it has an improper distribution with zero mean and an arbitrarily large variance matrix ($\boldsymbol{\Sigma}_\beta^{-1} \to 0$). The diffuse case captures the nonstationarity of a particular unobserved component and entails marginalising the inferences with respect to the parameter vector $\boldsymbol{\beta}$. As de Jong (1990) has shown, the posterior mean of $\boldsymbol{\beta}$ under the diffuse prior is coincident with the generalised least squares estimate of the parameter $\boldsymbol{\beta}$ considered as a fixed parameter vector in the classical sense. The only difference arises with respect to the definition of the likelihood.

6.3 Estimation of the factors and the disaggregated series

The factors and disaggregate values $\mathbf{Y}_{1t}$ are estimated by the following iterative scheme:

1. Start from a trial value $\mathbf{y}_{1t}^*, t = 0, \ldots, n$, (e.g. obtained from application of the univariate Chow-Lin disaggregation method, see Chow and Lin, 1971, to the first group of series, or the methodology in Moauro and Savio, 2005). In general, $\mathbf{y}_{1t}^*$ does not have to satisfy the temporal aggregation constraints.

2. Form the linear state space approximating model presented in (6.1) and (6.2), using the first-order Taylor expansion around $\mathbf{y}_{1t}^*$.

3. Use the Kalman filter and smoother to estimate the factors $\mathbf{f}_t$, the idiosyncratic components, and the disaggregate series $\mathbf{y}_{1t}$, and thus $\mathbf{Y}_{1t}$. In particular, if $\tilde{\boldsymbol{\alpha}}_{t|n}$ denotes the smoothed estimates of the state vector, the new estimate of the Box-Cox tranformation of the disaggregate series is obtained as

$$\hat{\mathbf{y}}_{1t}^* = [\mathbf{0}, \mathbf{I}, \mathbf{I}, \mathbf{0}]\tilde{\boldsymbol{\alpha}}_{t|n}.$$

4. If $||\mathbf{y}_{1t}^* - \hat{\mathbf{y}}_{1t}^*||$ is greater than a specified tolerance value, set $\mathbf{y}_{1t}^* = \hat{\mathbf{y}}_{1t}^*$ and return to step 2; else, set $\mathbf{Y}_{1t}^* = \mathbf{h}(\mathbf{y}_{1t}^*)$.

At convergence, the estimated disaggregate values satisfy the aggregation constraints, that is the observed quarterly aggregate $\mathbf{Y}_{1\tau}$ equals $\mathbf{h}(\mathbf{y}_{1,3\tau-1}^*) + \mathbf{h}(\mathbf{y}_{1,3\tau-2}^*) + \mathbf{h}(\mathbf{y}_{1,3\tau-3}^*)$. The relevant KFS for the linear approximating model is presented in the next section.

6.4 Univariate treatment of filtering and smoothing for multivariate models

The series $\mathbf{y}_t^{\dagger}$ is only partially observed and the KFS needs to be modified in order to entertain the missing values. Also, the state space form is formulated for the levels of the series in the first block, and thus the state vector has nonstationary effects. This section illustrates the KFS that is adapted to the state space model that takes into account the temporal aggregation observational constraints. The missing values affect systematically only the first block of N_1 elements of $\mathbf{y}_t^{\dagger}$: this situation can be dealt with if, for filtering purposes, the multivariate model is converted into a univariate model.

The univariate statistical treatment of multivariate models was considered by Anderson and Moore (1979). As we said before, it provides a very flexible and convenient device for filtering and smoothing in the presence of missing values. Our treatment is prevalently based on Koopman and Durbin (2000). However, for the treatment of initial conditions, and the estimation of $\boldsymbol{\beta}$, we adopt the augmentation approach by de Jong (1990).

The multivariate vectors $\mathbf{y}_t^{\dagger}$, $t = 1, \dots, n$, some elements of which can be missing, are stacked one on top of the other to yield a univariate time series $\{y_{t,i}^{\dagger}, i = 1, \dots, N_t, t = 1, \dots, n\}$, whose elements are processed sequentially; N_t is the number of time series processed at time t,

$$N_t = \begin{cases} N_1, & t = 0, \\ N_1 + N_2, & t = 1, 2, \dots, n. \end{cases}$$

The state space model for the univariate time series $\{y_{t,i}^{\dagger}\}$ is constructed as follows. The measurement equation for the i-th element of the vector $\mathbf{y}_t^{\dagger}$ is:

$$y_{t,i}^{\dagger} = \mathbf{z}_{t,i}^{'}\boldsymbol{\alpha}_{t,i} + c_{t,i} + \mathbf{g}_{t,i}^{\prime}\boldsymbol{\xi}_{2t}, \quad t = 0, \dots, n, \quad i = 1, \dots, N_t, \tag{8}$$

where $\mathbf{z}_{t,i}^{'}$ $\mathbf{g}_{t,i}^{'}$ and $c_{t,i}^{\prime}$ denote the i-th rows of $\mathbf{Z}_t$, $\mathbf{G}_t$ and $\mathbf{c}_t$, respectively.

The transition equation at time t varies according to i:

$$\boldsymbol{\alpha}_{t,i} = \begin{cases} \mathbf{T}_t\boldsymbol{\alpha}_{t-1,N_{t-1}} + \mathbf{d}_t + \mathbf{H}\boldsymbol{\eta}_{t,1}, & i = 1, \\ \boldsymbol{\alpha}_{t,i-1}, & i = 2, \dots, N_t. \end{cases}$$

The vector $\boldsymbol{\alpha}_{t,i}$ is the state vector when the (t,i)-th observation is processed. The state space form is completed by the initial state vector which is $\boldsymbol{\alpha}_{0,1} = \mathbf{a}_{0,0} + \mathbf{A}_{0,0}\boldsymbol{\beta} + \mathbf{H}_0\boldsymbol{\eta}_{0,0}$, where $\mathbf{P}_{0,0} = \text{Var}(\mathbf{H}_1\boldsymbol{\eta}_{1,1})$ and the other quantities have been defined in the previous section.

The augmented Kalman filter, taking into account the presence of missing values, is given by the following definitions and recursive formulae. Set the initial values $\mathbf{a}_{0,0} = \mathbf{d}_0, \mathbf{A}_{0,0}, \mathbf{P}_{0,0}, q_{0,0} = 0, \mathbf{s}_{0,0} = \mathbf{0}, \mathbf{S}_{0,0} = \mathbf{0}$, $cn = 0$; for $t = 0, \dots, n$, $i = 1, \dots, N_t - 1$, if $y_{t,i}^{\dagger}$ is available, compute the following quantities:

$$\begin{array}{lcllcl} v_{t,i} & = & y_{t,i}^{\dagger} - \mathbf{z}_{t,i}^{\prime}\mathbf{a}_{t,i} - c_{t,i}, & \mathbf{V}_{t,i}^{\prime} & = & -\mathbf{z}_{t,i}^{\prime}\mathbf{A}_{t,i}, \\ f_{t,i} & = & \mathbf{z}_{t,i}^{*'}\mathbf{P}_{t,i}\mathbf{z}_{t,i}^{*'} + \mathbf{g}_{t,i}^{\prime}\boldsymbol{\Psi}_2\mathbf{g}_{t,i}, & \mathbf{K}_{t,i} & = & \mathbf{P}_t\mathbf{z}_{t,i}^{*'}/f_{t,i} \\ \mathbf{a}_{t,i+1} & = & \mathbf{a}_{t,i} + \mathbf{K}_{t,i}v_{t,i}, & \mathbf{A}_{t,i+1} & = & \mathbf{A}_{t,i} + \mathbf{K}_{t,i}\mathbf{V}_{t,i}^{\prime}, \\ \mathbf{P}_{t,i+1} & = & \mathbf{P}_{t,i} - \mathbf{K}_{t,i}\mathbf{K}_{t,i}^{\prime}f_t, & & & \\ q_{t,i+1} & = & q_{t,i} + v_{t,i}^2/f_{t,i}, & \mathbf{s}_{t,i+1} & = & \mathbf{s}_{t,i} + \mathbf{V}_{t,i}v_{t,i}/f_{t,i}, \\ \mathbf{S}_{t,i+1} & = & \mathbf{S}_{t,i} + \mathbf{V}_{t,i}\mathbf{V}_{t,i}^{\prime}/f_{t,i} & d_{t,i+1} & = & d_{t,i} + \ln f_{t,i}, \\ cn & = & cn + 1 & & & \end{array} \tag{9}$$

Else, if $y^{\dagger}_{t,i}$ is missing, as it occurs for $\mathbf{Y}^c_{1t}$ for $t \neq 3\tau - 1, \tau = 1, \ldots, [(n+1)/3]$:

$$\begin{array}{rcl} \mathbf{a}_{t,i+1} & = & \mathbf{a}_{t,i}, \quad \mathbf{A}_{t,i+1} = \mathbf{A}_{t,i}, \\ \mathbf{P}_{t,i+1} & = & \mathbf{P}_{t,i}, \\ q_{t,i+1} & = & q_{t,i}, \quad \mathbf{s}_{t,i+1} = \mathbf{s}_{t,i}, \quad \mathbf{S}_{t,i+1} = \mathbf{S}_{t,i}, \quad d_{t,i+1} = d_{t,i}. \end{array} \tag{10}$$

Then for $i = N_t$

$$\begin{array}{rclrcl} \mathbf{a}_{t+1,1} & = & \mathbf{T}_{t+1}\mathbf{a}_{t,N_t} + \mathbf{d}_{t+1}, & \mathbf{A}_{t+1,1} & = & \mathbf{T}_{t+1}\mathbf{A}_{t,N_t}, \\ \mathbf{P}_{t+1,1} & = & \mathbf{T}_{t+1}\mathbf{P}_{t,N_t}\mathbf{T}'_{t+1} + \mathbf{H}\boldsymbol{\Sigma}_\eta\mathbf{H}', & & & \\ q_{t+1,1} & = & q_{t,N_t}, \quad \mathbf{s}_{t+1,1} = \mathbf{s}_{t,N_t}, & \mathbf{S}_{t+1,1} & = & \mathbf{S}_{t,N_t}, \quad d_{t+1,1} = d_{t,N_t}. \end{array} \tag{11}$$

Here, $\mathbf{V}_{t,i}$ is a vector with N_1 elements, $\mathbf{A}_{t,i}$ is $m \times (N_1)$, cn is the observation counter. The quantities $\mathbf{s}_{t,i}, \mathbf{S}_{t,i}$, accumulate vector and matrix cross-product that are used to build up the generalised least squares estimate of $\boldsymbol{\beta} = \mathbf{y}_{10}$. If the initial values are taken as fixed, maximising the likelihood with respect to $\boldsymbol{\beta}$ yields:

$$\hat{\boldsymbol{\beta}} = -\mathbf{S}^{-1}_{n+1,1}\mathbf{s}_{n+1,1}, \mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{S}^{-1}_{n+1,1}, \tag{12}$$

The profile log-likelihood is (neglecting constant terms)

$$\mathcal{L}_c = -\frac{1}{2}\left[d_{n+1,1} + q_{n+1,1} - \mathbf{s}'_{n+1,1}\mathbf{S}^{-1}_{n+1,1}\mathbf{s}_{n+1,1}\right]. \tag{13}$$

When $\boldsymbol{\beta}$ is diffuse (de Jong, 1991), the diffuse profile likelihood, denoted $\mathcal{L}_\infty$, takes the expression:

$$\mathcal{L}_\infty = -0.5\left[d_{n+1,1} + q_{n+1,1} - \mathbf{s}'_{n+1,1}\mathbf{S}^{-1}_{n+1,1}\mathbf{s}_{n+1,1} + \ln|\mathbf{S}_{n+1,1}|\right]. \tag{14}$$

Diagnostics and goodness of fit are based on the innovations, that are given by $\tilde{v}_{t,i} = v_{t,i} - \mathbf{V}'_{t,i}\mathbf{S}^{-1}_{t,i}\mathbf{s}_{t,i}$, with variance $\tilde{f}_{t,i} = f_{t,i} + \mathbf{V}'_{t,i}\mathbf{S}^{-1}_{t,i}\mathbf{V}_{t,i}$. The standardised innovations, $\tilde{v}_{t,i}/\sqrt{\tilde{f}_{t,i}}$ can be used to check for residual autocorrelation and departure from the normality assumption. The innovations have the following interpretation:

$$\tilde{v}_{t,i} = y^{\dagger}_{t,i} - \mathrm{E}(y^{\dagger}_{t,i}|\mathbf{Y}^{\dagger}_{t-1}, y^{\dagger}_{t,j}, j < i),$$

where $\mathbf{Y}^{\dagger}_t$ denotes the information set $\{\mathbf{y}^{\dagger}_1, \ldots, \mathbf{y}^{\dagger}_t\}$.

The filtered, or real time, estimates of the state vector and the estimation error matrix are computed as follows:

$$\tilde{\boldsymbol{\alpha}}_{t,i} = \mathbf{a}_{t,i} - \mathbf{A}_{t,i}\mathbf{S}^{-1}_{t,i}\mathbf{s}_{t,i} + \mathbf{P}_{t,i}\mathbf{z}_{t,i}\tilde{v}_{t,i}/f_{t,i}, \quad \tilde{\mathbf{P}}_{t,i} = \mathbf{P}_{t,i} + \mathbf{A}_{t,i}\mathbf{S}^{-1}_{t,i}\mathbf{A}^{*'}_{t,i} - \mathbf{P}_{t,i}\mathbf{z}_{t,i}\mathbf{z}'_{t,i}\mathbf{P}_{t,i}/f_{t,i},$$

where $\tilde{\boldsymbol{\alpha}}_{t,i} = \mathrm{E}(\boldsymbol{\alpha}_t|\mathbf{Y}^{\dagger}_{t-1}, y^{\dagger}_{t,j}, j \leq i)$, $\tilde{\mathbf{P}}_{t,i} = \mathrm{Var}(\boldsymbol{\alpha}_t|\mathbf{Y}^{\dagger}_{t-1}, y^{\dagger}_{t,j}, j \leq i)$.

The smoothed estimates are obtained from the augmented smoothing algorithm proposed by de Jong (1988), appropriately adapted to handle missing values. Defining $\mathbf{r}_{n,N} = 0, \mathbf{R}_{n,N} = 0, \mathbf{N}_{n,N} = 0$, for $t = n, \ldots, 0$, and $i = N_t, \ldots, 1$ if $y^{\dagger}_{t,i}$ is available:

$$\begin{array}{rcl} \mathbf{L}_{t,i} & = & \mathbf{I}_m - \mathbf{K}_{t,i}z^{*'}_i \\ \mathbf{r}_{t,i-1} & = & \mathbf{z}^*_{t,i}v_{t,i}/f_{t,i} + \mathbf{L}_{t,i}\mathbf{r}_{t,i}, \quad \mathbf{R}_{t,i-1} = \mathbf{z}^*_{t,i}\mathbf{V}'_{t,i}/f_{t,i} + \mathbf{L}_{t,i}\mathbf{R}_{t,i}, \\ \mathbf{N}_{t,i-1} & = & \mathbf{z}^*_{t,i}\mathbf{z}^{*'}_{t,i}/f_{t,i} + \mathbf{L}_{t,i}\mathbf{N}_{t,i}\mathbf{L}'_{t,i}. \end{array}$$

Else, if $y^{\dagger}_{t,i}$ is missing,

$$\mathbf{r}_{t,i-1} = \mathbf{r}_{t,i}, \quad \mathbf{R}_{t,i-1} = \mathbf{R}_{t,i}, \quad \mathbf{N}_{t,i-1} = \mathbf{N}_{t,i}.$$

$$\mathbf{r}_{t-1,N} = \mathbf{T}^{*'}_{t+1}\mathbf{r}_{t,i}, \quad \mathbf{R}_{t,i-1} = \mathbf{T}^{*'}_{t+1}\mathbf{R}_{t,i}, \quad \mathbf{N}_{t,i-1} = \mathbf{T}^{*'}_{t+1}\mathbf{N}_{t,i}\mathbf{T}^{*}_{t+1}.$$

The smoothed estimates of the state vector, $\tilde{\boldsymbol{\alpha}}_{t|n}$, along with their covariance matrices, $\mathbf{P}_{t|n}$, are obtained from the previously computed quantities as:

$$\begin{array}{rcl} \tilde{\boldsymbol{\alpha}}_{t|n} & = & \mathbf{a}_{t,1} + \mathbf{A}_{t,1}\tilde{\boldsymbol{\beta}} + \mathbf{P}_{t,1}(\mathbf{r}_{t-1,N} + \mathbf{R}_{t-1,N}\tilde{\boldsymbol{\beta}}) \\ \mathbf{P}_{t|n} & = & \mathbf{P}_{t,1} + \mathbf{A}_{t,1}\mathbf{S}^{-1}_{n+1}\mathbf{A}^{*'}_{t,1} - \mathbf{P}_{t,1}\mathbf{N}_{t-1,N}\mathbf{P}_{t,1}. \end{array}$$

From the smoothed estimates we obtain a new estimate of the disaggregate series on the Box-Cox transformed scale, $\mathbf{y}^{*}_{1,t}$, by computing $[\mathbf{0}, \mathbf{I}_{N1}, \mathbf{I}_{N1}, \mathbf{0}]'\tilde{\boldsymbol{\alpha}}_{t|n}$, and $\mathbf{Y}^{*}_{1,t} = \mathbf{h}(\mathbf{y}^{*}_{1,t})$.

7 Chain-linking and contemporaneous aggregation constraints

The quarterly national accounts series are subject to a number of accounting deterministic constraints, when the aggregates are expressed at current prices and at the average prices of the previous year. In particular, the 17 series listed in 1 are bound together by the identities:

GDP at basic prices	=	$\sum$ Value added of the 6 branches (A-B, C-D-E, F, G-H-I, J-K, L-P)
GDP at market prices	=	GDP at basic prices + Taxes less subsidies
GDP at market prices+IMP	=	CONS+INV+EXP
Domestic demand	=	CONS+INV
CONS	=	CONS$_H$+CONS$_G$

where

CONS	=	Final consumption expenditures
CONS$_H$	=	Household and NPISH final consumption expenditure
CONS$_G$	=	Final consumption expenditure: general government
INV	=	Gross Capital Formation
EXP	=	Exports of goods and services
IMP	=	Imports of goods and services

The production of chained linked national accounts estimates has changed drastically the role of the contemporaneous aggregation constraints considered above. In particular, the constraints hold only when the series are expressed at the average prices of the previous year; loosely speaking, only in that case they are expressed genuinely at constant prices. Otherwise, chaining, which is a multiplicative operation, destroys the additivity of the constraints, and a nonzero discrepancy arises. GDP and its main components are expressed in chain-linked volumes (millions of euros), with reference year 2000, which implies that the constraints hold exactly for the four quarters of the year 2001. Interestingly, due to the application of the *annual overlap technique*, exposed below, the constraints are not entirely lost, but they continue to hold after a transformation of the data that we call "dechaining", which aims at expressing the chained values at the prices of the previous year.

The Eurozone member states chain-link the quarterly data on an annual basis, i.e. the quarterly volume measures are expressed at the average prices of the previous years. The current situation is described in the

Eurostat metadata available at http://ec.europa.eu/eurostat. Two alternative techniques are applied for annual chain-linking of quarterly data by the member countries: one quarter overlaps (Austria) and annual overlaps (other states). These are described in Bloem, *et al.* (2001, chapter IX); the annual overlap technique, which implies compiling estimates for each quarter at the weighted annual average prices of the previous year, has the the advantage of producing quarterly volume estimates that add up exactly to the corresponding annual aggregate. The annual overlap technique is also the method used by Eurostat in the imputation of the chain-linked volume measures of those countries for which no quarterly data at previous years prices are available.

As it is well known, chain-linking results in the loss of cross-sectional additivity (if the one quarter overlap is used also temporal additivity is lost and benchmarking techniques have to be employed in order to restore it). However, for the annual overlap, the disaggregated (monthly and quarterly) volume measures expressed at the prices of the previous year preserve both the temporal and cross-sectional additivity.

The cross-sectional constraints can be enforced by a multistep procedure that de-chains the estimated monthly values, expressing them at the average prices of the previous year, and projects the estimates on the subspace of the constraints, as it will be described below. The dechaining procedure is in line with that advocated by the IMF manual (see Bloem *et al.*, 2001).

We start by indexing the month of the year by $j, j = 0, \ldots, 11$ and the year by m, $m = 1, \ldots, M = [(n+1)/12]$, so that the time index is written $t = j + 12m, t = 0, \ldots, n$.

For a particular estimated monthly time series let us denote by Y_{jm} the value at current prices of month j in year m, $Y_{.m} = \sum_j Y_{jm}$ the annual total, $\bar{Y}_m = Y_{.m}/12$ the annual average (the annual and quarterly figures are available from the national accounts, compiled by Eurostat). The chain-linked volume estimate with reference year b (the year 2000 in our case) will be denoted $\hat{Y}_{jm}^{(b)}$. The temporal disaggregation methods described in the previous section are applied to the quarterly chained-linked volume series with reference year b and yield estimates that add up to the quarterly and annual totals (temporal consistency), but are not additive in a horizontal (that is cross-sectional) sense.

The following multistep procedure enables the computation of volume measures expressed at the prices of the previous year that are additive, also horizontally.

1. *Dechaining* :

   (a) Transform the monthly estimates into Laspayres type quantity indices with reference year b (volumes are evaluated at year b average prices), by computing

   $$I_{jm}^{(b)} = \frac{\hat{Y}_{jm}^{(b)}}{\bar{Y}_{.b}}, j = 0, \ldots, 11, m = 0, \ldots, M,$$

   where the denominator is the annual average of year b at current prices. In our case $b = 5$ (year 5 is the calendar year 2000).

   (b) Change the reference year to $m = 1$, the second year of the series (1996 in our case), by computing:

   $$I_{jm}^{(1)} = \frac{I_{jm}^{(b)}}{\bar{I}_1^{(b)}}, j = 1, \ldots, 11, m = 0, \ldots, M,$$

   where $\bar{I}_1^{(b)} = \sum_j I_{j1}^{(b)}/12$ is the average quantity index for the second year of the the sample.

(c) Transform the quantity indices for year $m = 1, 2, \ldots, M$ into indices with reference year $m-1$ (the previous year), by rescaling $I_{jm}^{(1)}$ as follows:

$$I_{jm}^{(m-1)} = \frac{I_{jm}^{(1)}}{\bar{I}_{m-1}^{(1)}}, j = 0, \ldots, 11, m = 1, \ldots, M,$$

where

$$\bar{I}_{m-1}^{(1)} = \frac{1}{12}\sum_j I_{j,m-1}^{(1)}, m = 1, \ldots, M$$

(d) Compute the series at the average prices of the previous year as:

$$\hat{Y}_{jm}^{(m-1)} = I_{jm}^{(m-1)}\bar{Y}_{m-1}, \quad j = 0, \ldots, 11, m = 1, \ldots, M,$$

2. *Aggregation step*: Let $\mathbf{Y}_t^{(m-1)}$ denote the disaggregate time series expressed at the average prices of the previous year. Using the original estimates and the dechaining procedure we can assume that, at least approximately,

$$\mathbf{Y}_t^{(m-1)} \sim \mathbf{N}\left(\hat{\mathbf{Y}}_t^{(m-1)}, \hat{\mathbf{V}}_t^{(m-1)}\right), t = 0, 1, \ldots, n,$$

where the first and second moments are given by the sequential constrained estimates produced by the Kalman filter and smoother outlined in the previous section, modified to take into account the dechaining procedure[1]. If the r cross-sectional constraints are expressed as

$$\mathbf{Q}\mathbf{Y}_t = \mathbf{q}$$

where $\mathbf{Q}$ is an $r \times N_1$ matrix, and $\mathbf{q}$ is $r \times 1$, the modified estimates that comply with those constraints and their MSE matrix are given respectively by

$$\begin{aligned}
\tilde{\mathbf{Y}}_t^{(m-1)} &= \hat{\mathbf{Y}}_t^{(m-1)} + \hat{\mathbf{V}}_t^{(m-1)}\mathbf{Q}'(\mathbf{Q}\hat{\mathbf{V}}_t^{(m-1)}\mathbf{Q}')^{-1}(\mathbf{q} - \mathbf{Q}\hat{\mathbf{Y}}_t^{(m-1)}) \\
\tilde{\mathbf{V}}_t^{(m-1)} &= \hat{\mathbf{V}}_t^{(m-1)} - \hat{\mathbf{V}}_t^{(m-1)}\mathbf{Q}'(\mathbf{Q}\hat{\mathbf{V}}_t^{(m-1)}\mathbf{Q}')^{-1}\mathbf{Q}\hat{\mathbf{V}}_t^{(m-1)}
\end{aligned}$$

see, e.g. Peña (1997). In our case, $r = 5$ and $\mathbf{q} = \mathbf{0}$,

$$\mathbf{Q} = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1
\end{bmatrix}, \mathbf{q} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

The new balanced estimates are now ready to be expressed at the average prices of reference year b.

3. *Chain-linking* (annual overlap):

[1]In particular, if $\mathcal{D}_t$ is a diagonal matrix containing the dechaining coefficients that allow to express the chained estimates at the average prices of the previous year, and $\hat{\mathbf{Y}}_t$ are the chained estimates, with estimation error mean square matrix $\hat{\mathbf{V}}_t$, computed by the Kalman filter and smoother, then $\hat{\mathbf{Y}}_t^{(m-1)} = \mathcal{D}_t\hat{\mathbf{Y}}_t$ and $\hat{\mathbf{V}}_t = \mathcal{D}_t\hat{\mathbf{V}}_t\mathcal{D}_t$.

(a) Convert the aggregated volume measures into Laspeyres-type quantity indices with respect to the previous year:

$$\mathcal{I}_{jm}^{(m-1)} = \frac{\tilde{Y}_{jm}^{(m-1)}}{\bar{Y}_{m-1}}, \quad j = 0, \dots, 11, m = 1, \dots, M,$$

where $\bar{Y}_{m-1} = \sum_j Y_{j,m-1}/12$ is the average of the previous year at current prices. The annual and quarterly totals is available from the national accounts compiled by Eurostat.

(b) Chain-link the indices using the recursive formula (the first year is the reference year):

$$\mathcal{I}_{jm}^{(0)} = \mathcal{I}_{jm}^{(m-1)} \bar{\mathcal{I}}_{m-1}^{(0)}, \quad j = 0, \dots, 11, m = 1, \dots, M,$$

where $\bar{\mathcal{I}}_0^{(0)} = 1$ and

$$\bar{\mathcal{I}}_{m-1}^{(0)} = \frac{1}{12} \sum_j \mathcal{I}_{j,m-1}^{(0)}.$$

(c) If $b > 0$ then change the reference year to year b:

$$\mathcal{I}_{jm}^{(b)} = \frac{\mathcal{I}_{jm}^{(0)}}{\bar{\mathcal{I}}_b^{(0)}} \quad j = 0, \dots, 11, m = 1, \dots, M.$$

(d) Compute the chain-linked volume series with reference year b:

$$\tilde{Y}_{jm}^{(b)} = \mathcal{I}_{jm}^{(b)} \bar{Y}_b \quad j = 1, \dots, 12, m = 2, \dots, M,$$

where $\bar{Y}_b = \frac{1}{12} \sum_j Y_{jb}$ is the value of GDP (at basic or market prices) at current prices of the reference year.

The multistep procedure just described enables to obtain monthly estimates in volume such that the values $\tilde{Y}_{jm}^{(m-1)}$ expressed at the average prices of the previous year add up to their quarterly and annual totals published by Eurostat and are consistent with the contemporaneous aggregation constraints. On the contrary, as a result of the chaining procedure, the chain-linked volumes $\tilde{Y}_{jm}^{(b)}$ expressed at the prices of the common reference year b (2000) are consistent only with the temporal aggregation constraints; however, their estimates are more reliable since they have been combined with the estimates of other related variables.

8 Estimation Results

We now put the pieces together and estimate the factor model using the dataset consisting of 149 time series with mixed frequency described in section 2. For all the series considered in our application, we can assume that the monthly logarithmic changes are stationary, so that $\lambda_i = 1$ and $\varphi_i = 1$, except for the Business and Consumer Survey series, for which $\lambda_i = 0$ and $\varphi_i = 0$. In particular, the survey variables will require no transformation as they are expressed as balances. All the remaining series measured on a ratio scale are transformed into logarithms.

Estimation of the unknown parameters and temporal disaggregation is carried out by an iterative algorithm which alternates two main steps until convergence. We start from a trial disaggregate time series

$\mathbf{y}^*_{1t}, t = 0, \ldots, n,$, obtained from the temporal disaggregation of the quarterly national accounts series according to the univariate Chow-Lin procedure, using industrial production and retail sales (total) as monthly indicators (see Proietti, 2007). The disaggregate time series serve to construct the standardized stationary series $\mathbf{x}_t$, that form a balanced panel of monthly time series. The initial estimate of the parameter is computed by a principal component analysis of the covariance matrix of the $\mathbf{x}_t$'s.

The number of factors, K, is selected at this stage according to the information criteria proposed by Bai and Ng (2002). In particular we focus on the two criteria:

$$IC_{p1}(k) = \ln V(k) + k\left(\frac{n+N}{nN}\right)\ln\left(\frac{nN}{N+T}\right),$$

$$IC_{p2}(k) = \ln V(k) + k\left(\frac{n+N}{nN}\right)\ln\min\{n, N\},$$

where $V(k) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{n}(x_{it} - \hat{\theta}_i\hat{f}_{it})^2$, and the factors are estimated by static principal components analysis. Bai and Ng (2002) show that the value of k that minimizes $IC_{p1}(k)$ or $IC_{p1}(k)$ is a consistent estimator for $n, N \to \infty$ of the number of common factors.

Conditional on K, the estimation of the factor model involves the following steps:

1. Given a set of estimated disaggregate values $\hat{\mathbf{y}}_{1t}$, satisfying the temporal and contemporaneous aggregation constraints, we construct the pseudo complete balanced panel of time series $\mathbf{y}_t = [\hat{\mathbf{y}}'_{1t}, \mathbf{y}'_{2t}]'$, where $\mathbf{y}_{2t}$ are the observed monthly series. We then obtain the stationary transformation $\boldsymbol{\Delta}\mathbf{y}_t$ and estimate $\boldsymbol{\mu}$ and $\mathbf{D}$ by computing the sample average and the standard deviation of the individual time series. We construct the standardized stationary series $\mathbf{x}_t = \hat{\mathbf{D}}^{-1}(\boldsymbol{\Delta}\mathbf{y} - \hat{\boldsymbol{\mu}})$, and estimate the parameters of the factor model $\boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}_\eta, \boldsymbol{\Psi}$ by maximum likelihood using the EM algorithm (see section 4.1) or by principal component analysis (see section 4.2).

2. Conditional on the parameter estimates, we estimate the disaggregate time series $\hat{\mathbf{y}}_{1t}$ (and thus $\hat{\mathbf{Y}}_{1t} = \mathbf{h}(\hat{\mathbf{y}}_{1t})$), consistent with the temporal and cross-sectional constraints. This step is carried out iteratively, with each iteration consisting of two steps:

   (a) estimate $\hat{\mathbf{y}}_{1t}$ enforcing the nonlinear temporal aggregation constraints, as detailed in (6.3);

   (b) enforce the cross-sectional temporal aggregation constraints by the de-chaining and chaining-back procedure outlined in section (7).

Convergence occurs when both the parameters $\boldsymbol{\Xi}$ and the estimates of the disaggregate time series $\hat{\mathbf{y}}_{1t}$ do not differ from one iteration to another by more than a specified tolerance (10^{-5}).

The estimated number of factors is $K = 6$: this can be considered as a conservative estimate. The plot of the Bai and Ng (2002) information criteria, presented in the first panel of figure 1, reveals that the $IC_{p1}(k)$ suggests the choice of 6 common factors, whereas $IC_{p2}(k)$ has its minimum at $K = 3$. The share of the variance explained by the first three principal components is 34.13%, whereas that explained by the first six is 45.18%.

The estimation of the factor model was carried out using both the EM algorithm and PCA, as far as the estimation of the parameter vector $\boldsymbol{\Xi}$ is concerned. Less than 200 iterations are required for convergence in both cases. The right panel of figure 1 displays the value of the likelihood (14) versus the iteration number for the two methods. The estimation results are very similar, both for $\boldsymbol{\Xi}$ and the disaggregate series; however, the estimated factors conditional on the PCA parameter estimates are slightly smoother than those

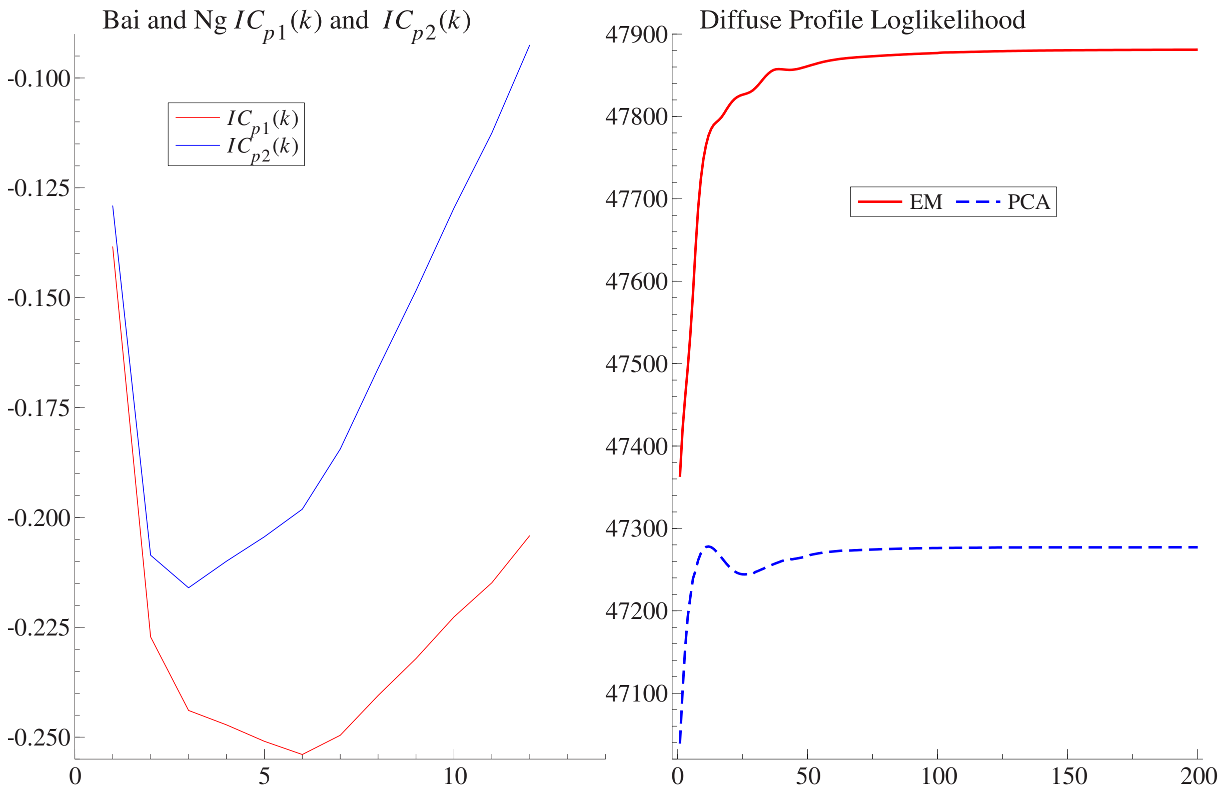

Figure 1: Estimation of the number of common factors: Bai and Ng Information Criteria against the number of factors (left panel). Convergence of the EM and PCA estimation methods (right panel).

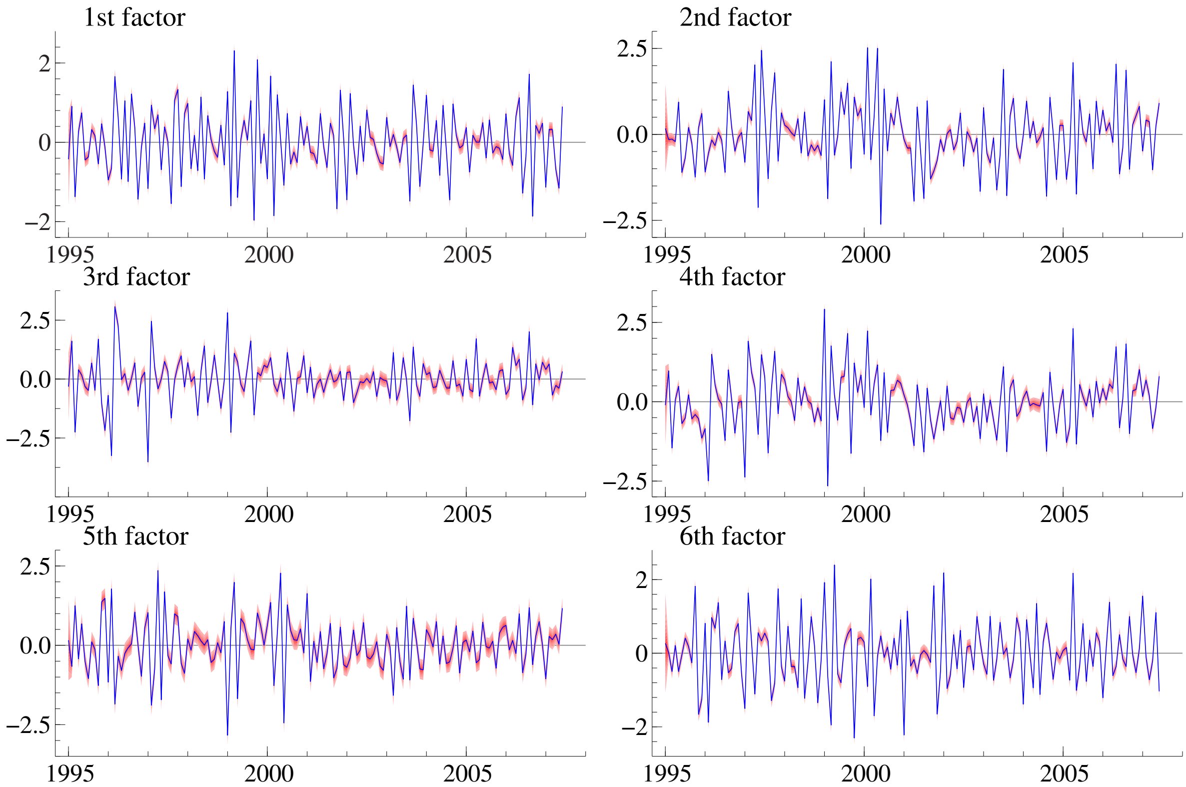

Figure 2: Point and 95% interval estimates of the common factors.

obtained from the EM method. As a consequence, the disaggregate series $\hat{\mathbf{Y}}_{1t}$ have a smaller variation at the high frequencies. Since *ceteris paribus* we would prefer smoother estimates of monthly GDP and its components, the presentation of the results will henceforth concentrate on the PCA method. It should be recalled that PCA is used only for estimating the parameters in Ξ; the factors are estimated along with the monthly GDP and its components according to the second step of our procedure (i.e. incorporating the temporal and cross-sectional aggregation constraints).

Figure 2 displays the point estimates of the six factors, $\tilde{\mathbf{f}}_{t|n}$, and the approximate 95% interval estimates, based on the assumption of normality. As the plot illustrates, the dynamic of the estimated factors is dominated by high frequency variation, resulting in a negative autocorrelation; also, the third factor captures the main economic shocks that affected the construction sectors. However, the factors capture also the dynamics of the euro area business cycle: in particular, this information is carried by the 2nd, 4th and 5th factors, as can be seen from figure 3, which shows the Baxter and King (1999) cyclical component in the estimated factors.

Figure 4 is a biplot of the estimated factor analysis (see Gower and Hand, 1996). Letting $\hat{\boldsymbol{\Theta}} = \mathcal{U}\mathcal{M}\mathcal{V}'$ denote the singular value decomposition of the loadings matrix, each individual series is represented by a point in the plot with coordinates provided by the first two columns of $\mathcal{U}\mathcal{M}$, whereas the factors are represented by lines from the origin, with coordinates given by the first two columns of $\mathcal{V}$. The length of the line drawn from the origin is an approximation of the variance of the columns of the loadings associated

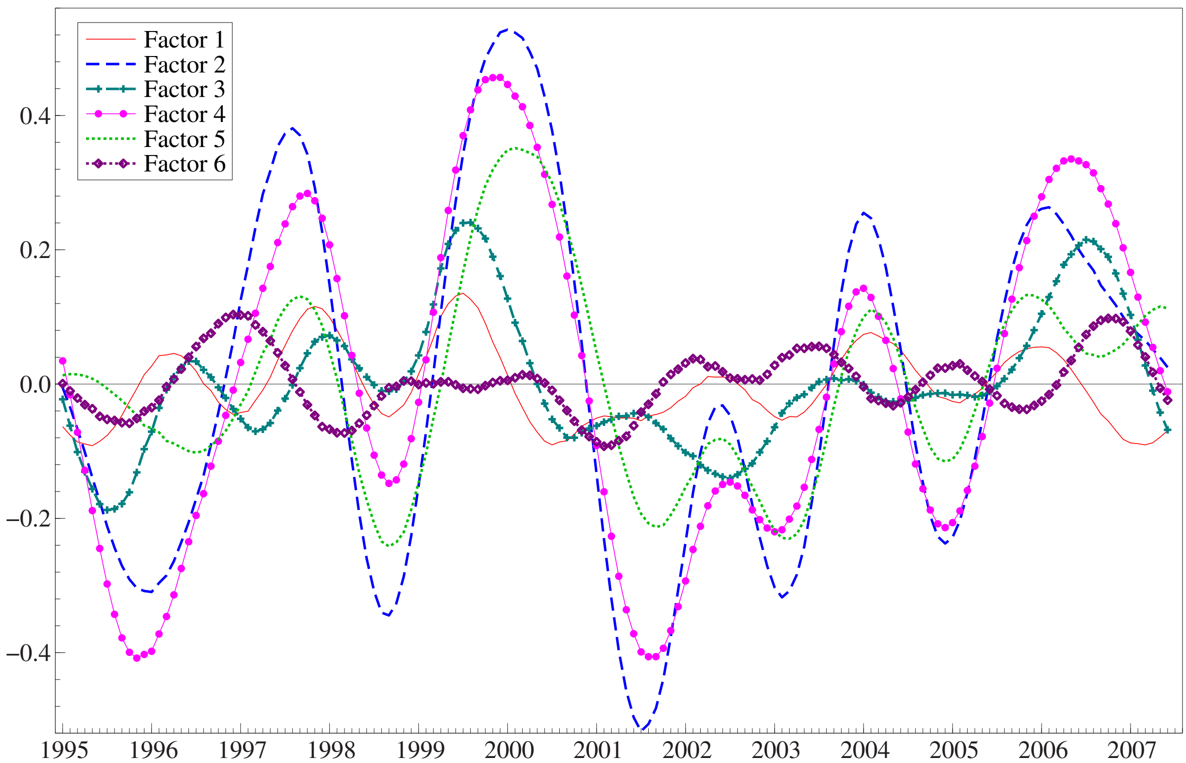

Figure 3: Baxter and King cyclical components of the factors.

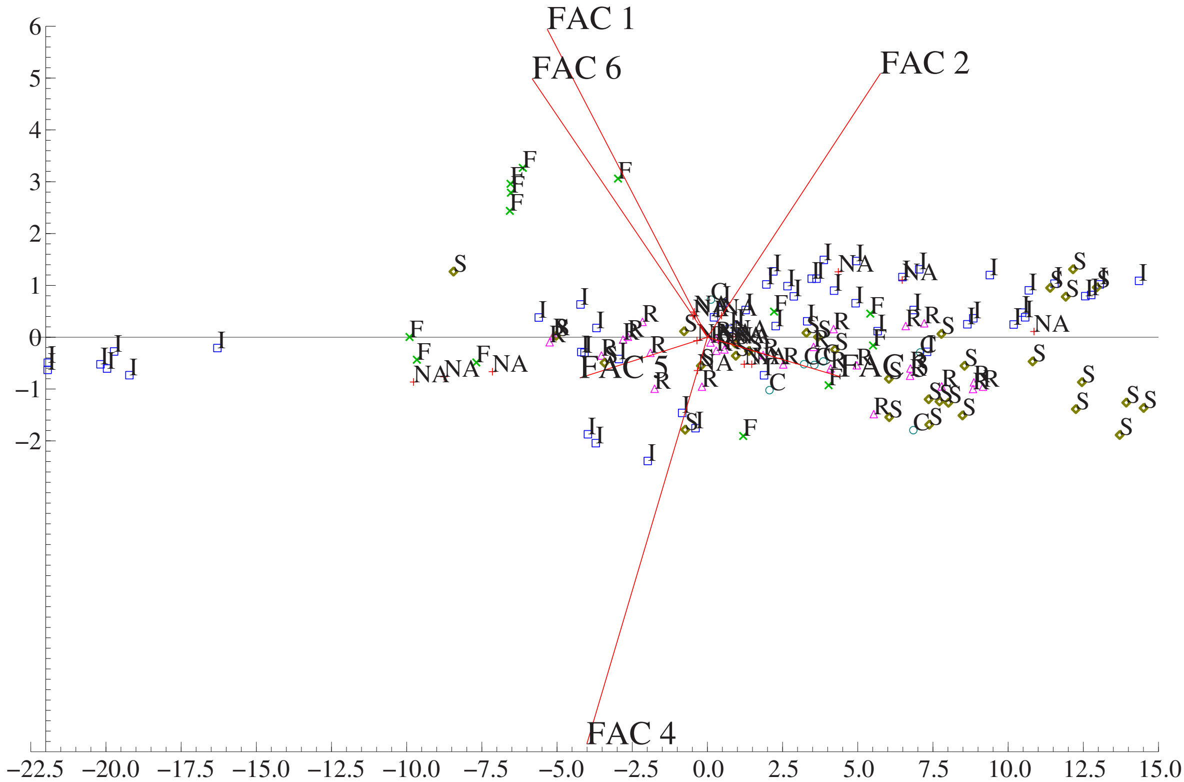

Figure 4: Biplot of the factor loadings.

to a particular factor; the cosine of the angle formed by any two lines is an approximation of the correlation between two columns of $\hat{\Theta}$. Series that load on the same factors will be represented by two close points; the labels "NA", "I", "C", "R", "F", "S" refer, respectively, to the national accounts series, industry, construction, retail, financial and monetary indicators, business and consumer surveys. A group of series with the same loadings pattern is hours worked in industry, displayed to the left of the biplot. In general, series belonging to the same group tend to cluster together. The loading of a particular variable on a specific factor can be approximated by the orthogonal projection of the point representing the variable on the line representing the factor. The survey series are mostly related to the second and the third factors, whereas the financial variables are associated to the 4th and 6th factors. The monthly construction series are mostly associated to factor 3 (the loading of value added in the construction sector is 1).

A most important side output of our modeling effort is the estimation of monthly GDP and its main components. The estimates comply with the temporal aggregation constraints and the cross-sectional identities for the year 2001, and if the series are expressed at the prices of the previous year. Moreover, they are highly informative as they incorporate the information that is common to a large set of monthly indicators. Figure 5 displays monthly GDP at market prices, final consumption expenditures and gross capital formation, along with their monthly and yearly growth rates. It must be stressed that approximate measures of reliability of the estimates are directly available from the our methodology. The interval estimates, also presented in figure 5, reveal that the estimation error variance is lower for GDP than for gross capital formation, and that

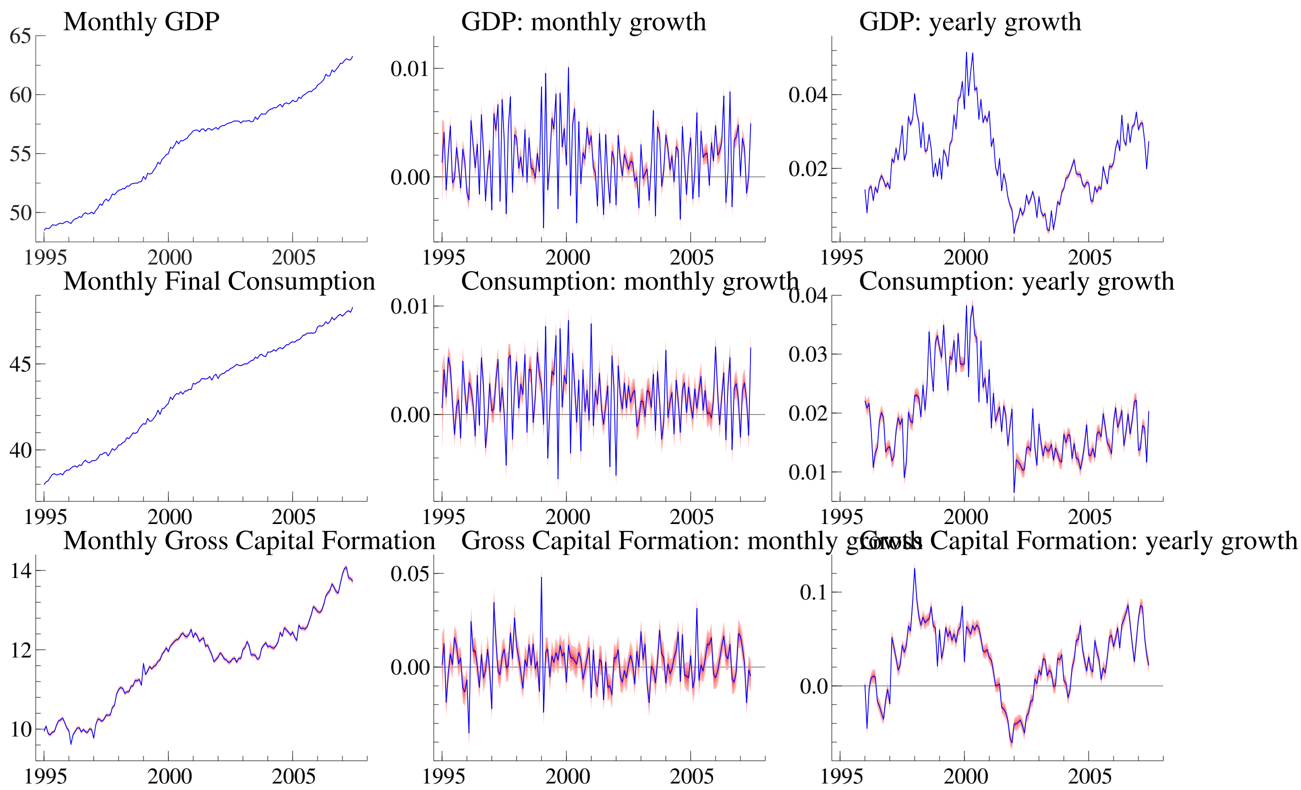

Figure 5: Monthly estimates of GDP at market prices, Final Consumption and Gross Capital Formation (chained 2000 volumes), and monthly and yearly growth. Point and 95% interval estimates.

it is generally low and, as our experience suggests, lower than that obtained from other univariate and small scale multivariate models (see Proietti, 2007, and Proietti and Moauro, 2006). Finally, figure 6 displays the estimates of value added for the six branches of the NACE classification, that together make up GDP at basic prices, and their yearly growth.

9 Conclusive remarks

The paper has proposed an iterative scheme for estimating a large scale factor model with data at different frequencies, providing an exact treatment of the temporal and cross-sectional aggregation constraints. The model is used to nowcast monthly GDP and its decomposition by expenditure type and by the output approach. The results are relevant not only because the estimated common factors embody the economic information contained in the national accounts macro variables, but also because the availability of monthly estimates of the national accounts series can be seen as a useful addition to the available published data.

There are three important points that we did not address in the paper and that we leave to our future research agenda: the first is to carry out a real time experiment to assess the process of updating the nowcasts of monthly GDP and its main components as new releases of data become available (see Giannone, Reichlin and Small, 2006). The second is to incorporate parameter uncertainty in the assessment of the reliability of

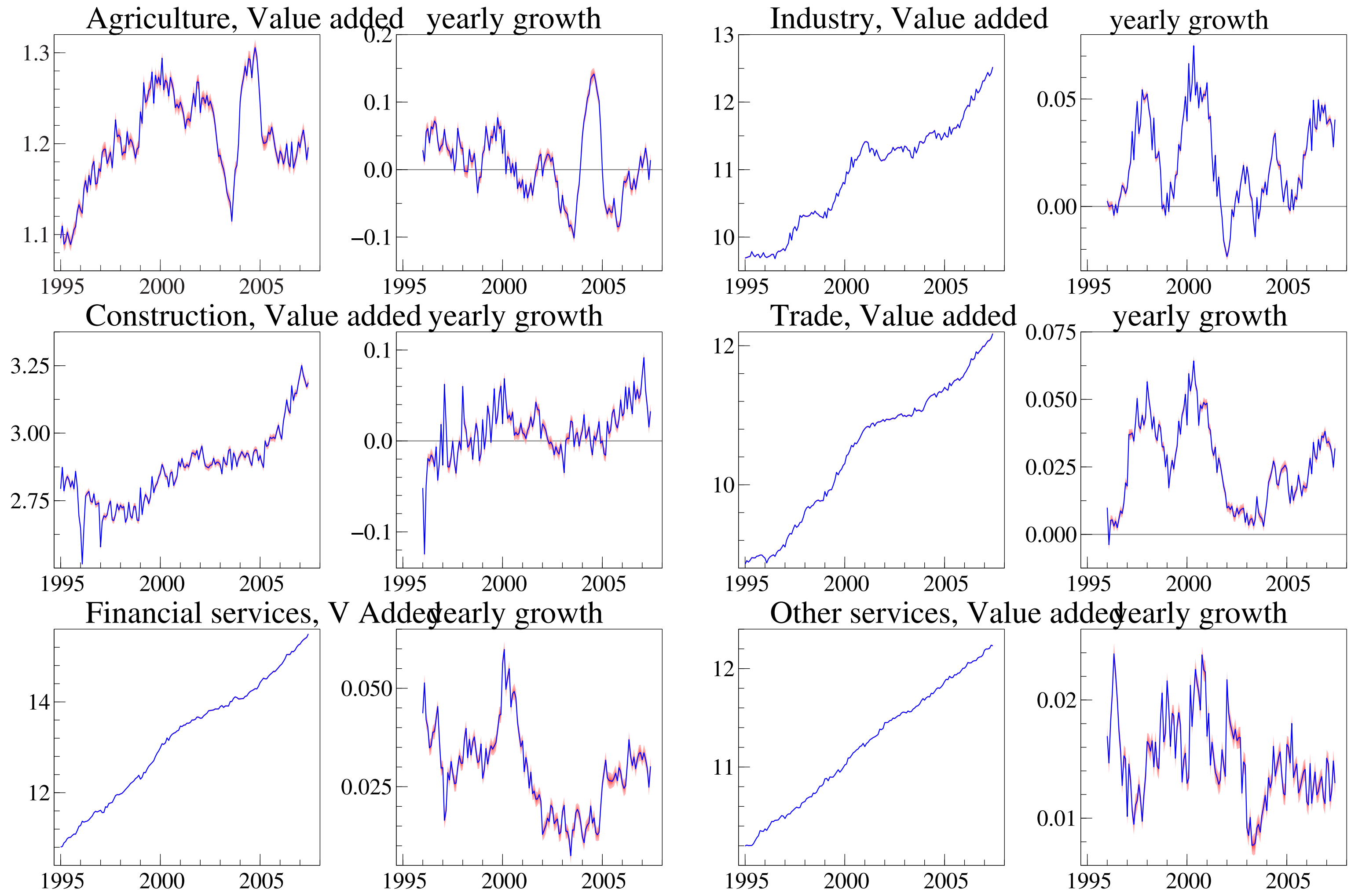

Figure 6: Monthly estimates of valued added for the six branches. Chained 2000 volumes and yearly growth, point and 95% interval estimates.

the nowcasts. In facts, the measures are reliability made available by the nonlinear smoothing algorithm that was proposed in the paper are conditional on the parameters estimates obtained by the last step of the EM algorithm or principal components. The final issue is that of nowcasting the low-pass component monthly GDP growth, as in Altissimo et al. (2006), within our model based framework. This could be carried out by embedding a decomposition of output fluctuations within the factor model, along the lines of Proietti and Musso (2007).

References

Altissimo, F., Cristadoro, R., Forni, M., Lippi, M., and Veronese, G. (2007). New Eurocoin: Tracking Economic Growth in Real Time. Temi di discussione (Economic working papers) 631, Bank of Italy, Economic Research Department.

Amenegual, D., and Watson, M.W. (2007). Consistent estimation of the number of dynamic factors in a large N and T panel. *Journal of Business and Economic Statistics*, 25, 91–96.

Anderson, B.D.O., and Moore, J.B. (1979). *Optimal Filtering*, Englewood Cliffs NJ: Prentice-Hall.

Angelini, E., Henry, J. and Marcellino, M. (2004). Interpolation and back-dating with a large information set. CEPR Discussion Paper Series 4533.

Aruoba, S.B., Diebold, F.X., and Scotti, C. (2007). Real-time measurement of business conditions, International Finance Discussion Papers 901, Board of Governors of the Federal Reserve System (U.S.).

Bai, J. (2003). Inferential theory for factor models of large dimension. *Econometrica*, 71, 135-71.

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

Bańbura, M., and Rünstler, G. (2007). A look into the factor model black box. Publication lags and the role of hard and soft data in forecasting GDP. ECB working paper No 751. European Central Bank, Frankfurt am Main.

Baxter, M., and King, R.G. (1999). Measuring Business Cycles: Approximate Band-Pass Filters for Economic Time Series. *The Review of Economics and Statistics*, 81, 575–593.

Bloem, A., Dippelsman, R.J. and Maehle, N.O. (2001), *Quarterly National Accounts Manual Concepts, Data Sources, and Compilation*, International Monetary Fund.

Boivin, J., Ng, S. (2004). Are more data always better for factor analysis? Journal of Econometrics, 132, 169–194.

Box G. E. P., and Cox, D.R. (1964). An analysis of transformations (with discussion), *Journal of the Royal Statistical Society, B*, 26, 211–246.

Breitung, J., and Schumacher, C. (2006). Real-time forecasting of GDP based on a large factor model with monthly and quarterly data. Deutsche Bundesbank Discussion Paper 33.

Bryson, A.E., and Ho, Y.C. (1969). *Applied optimal control: optimization, estimation, and control*. Blaisdell Publishing, Waltham, Mass.

Chow, G., and Lin, A. L. (1971). Best Linear Unbiased Interpolation, Distribution and Extrapolation of Time Series by Related Series, *The Review of Economics and Statistics*, 53, 4, 372-375.

Cristadoro, R., Forni, M., Reichlin, L. and Veronese, G. (2005). A core inflation indicator for the euro area. *Journal of Money Credit and Banking*, 37, 539–560.

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39(1):138.

de Jong, P. (1989). Smoothing and interpolation with the state space model, *Journal of the American Statistical Association*, 84, 1085-1088.

de Jong, P. (1991). The diffuse Kalman filter, *Annals of Statistics*, 19, 1073-1083.

de Jong, P., and Mackinnon, M.J. (1988). Covariances for smoothed estimates in state space models. *Biometrika*, 75, 601–602.

Doornik, J.A. (2006). *Ox 4.0 - An Object-Oriented Matrix Programming Language*, Timberlake Consultants Ltd: London.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data. *Journal of the Royal Statistical Society*, 14, 1:38.

Doz, C., Giannone, D., and Reichlin, L. (2005). A two-step estimator for large approximate dynamic factor models based on kalman filtering. CEPR Discussion Paper No. 6043.

Doz, C., Giannone, D., and Reichlin, L. (2007). A Quasi Maximum Likelihood Approach for Large Approximate Dynamic Factor Models, CEPR Discussion Paper No. 5724

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

Eurostat (1999), Handbook of quarterly national accounts, Luxembourg, European Commission.

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

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2001). Coincident and leading indicators for the euro area. *The Economic Journal*, 111, C62–C85.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The generalized factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association*, 100, 830-40.

Geweke, J.F., and Singleton, K.J. (1981). Maximum likelihood confirmatory factor analysis of economic time series. *International Economic Review*, 22, 1980.

Giannone, D., Reichlin, L., and Sala, L. (2005), Monetary Policy in Real Time, in NBER Macroeconomic Annual 2004, eds. M. Gertler and K. Rogoff, Cambridge, MA: MIT Press, pp. 161200.

Giannone, D., Reichlin L. and Small, D. (2006). Nowcasting GDP and Inflation: the real-time informational content of macroeconomic data releases. ECB working paper n. 633.

Gill, P.E., Murray, W., Saunders, M.A., and Wright, M.H. (1989). Constrained nonlinear programming, in G.L. Nemhauser, A.H.G. Rinnooy Kan, M.J. Todd (eds.), *Handbooks in Operations Research and Management Science*, Vol. 1, Optimization, pp. 171-210, Elsevier, Amsterdam, 1989.

Gower, J. C. and Hand, D. J. (1996) *Biplots*. London, Chapman & Hall.

Hallin, M., and R. Liska (2007). Determining the Number of Factors in the Generalized Factor Model, *Journal of the American Statistical Association*, 102, 603-617.

Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press: Cambridge.

Harvey, A.C. and Chung, C.H. (2000) Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistics Society*, Series A, 163, 303-339.

Koopman, S. J. (1993). Disturbance smoother for state space models. *Biometrika*, 80, 117-26.

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

Mariano, R.S., and Murasawa, Y. (2003). A new coincident index of business cycles based on monthly and quarterly series. *Journal of Applied Econometrics*, 18, 427-443.

Mitchell, J., Smith, R.J., Weale, M.R., Wright, S. and Salazar, E.L. (2005). An Indicator of Monthly GDP and an Early Estimate of Quarterly GDP Growth. *Economic Journal*, 115, 108–129.

Mönch, E., and Uhlig, H. (2004). Towards a Monthly Business Cycle Chronology for the Euro Area. CEPR Discussion Papers 4377.

Moauro F. and Savio G. (2005). Temporal Disaggregation Using Multivariate Structural Time Series Models. *Econometrics Journal*, 8, 214-234.

Peña, D. (1997). Combining Information in Statistical Modeling. *The American Statistician*, 51, 326–332.

Proietti, T. (2006). Temporal Disaggregation by State Space Methods: Dynamic Regression Methods Revisited. *Econometrics Journal*, Vol. 9, pp. 357-372.

Proietti, T. (2006). On the estimation of nonlinearly aggregated mixed models. *Journal of Computational and Graphical Statistics*, Vol. 15, 1–21.

Proietti T. and Moauro F. (2006). Dynamic Factor Analysis with Nonlinear Temporal Aggregation Constraints. *Journal of the Royal Statistical Society*, series C, 55, pp. 281300.

Proietti T., and Musso A. (2007). Growth accounting for the euro area: a structural approach. ECB Working Paper Series No. 804, European Central Bank, Frankfurt am Main.

Shumway, R.H., and Stoffer, D.S. (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 Watson M.W. (1991). A probability model of the coincident economic indicators. In *Leading Economic Indicators*, Lahiri K, Moore GH (eds), Cambridge University Press, New York.

Stock, J.H., and M.W. Watson (2002a). Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business and Economic Statistics*, 20, pp.147-162.

Stock, J.H., and M.W. Watson (2002b). Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association*, 97, 1167-79.

Stock, J.H., and M.W. Watson (2006). Forecasting with many predictors, in Elliott, G., Granger, C.W.J. and Timmermann, A. (eds.) *Handbook of Economic Forecasting, Volume 1*, Elsevier.

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

A List of the time series employed in the paper

Table 1: Quarterly time series: National accounts aggregates

Table	Description	Name
namq_nace06_k	Value added, Agriculture, hunting, forestry and fishing	a-b
namq_nace06_k	Value added, Industry, incl. Energy	c-d-e
namq_nace06_k	Value added, Construction	f
namq_nace06_k	Value added, Trade, transport and communication services	g-h-i
namq_nace06_k	Value added, Financial services and business activities	j-k
namq_nace06_k	Value added, Other services	c-d-e
namq_gdp_k	Gross domestic product at market prices	b1gm
namq_gdp_k	Final consumption expenditure	p3
namq_gdp_k	Domestic demand	p3_p5
namq_gdp_k	Household and NPISH final consumption expenditure	p31_s14_s15
namq_gdp_k	Final consumption expenditure: general government	p3_s13
namq_gdp_k	Gross capital formation	p5
namq_gdp_k	Gross fixed capital formation	p51
namq_gdp_k	Exports of goods and services	p6
namq_gdp_k	Imports of goods and services	p7
nama_q	Total Gross Value Added (at basic prices)	b1g
namq_gdp_k	Taxes less subsidies on production and imports	d21_m_d31

Table 2: Monthly time series: Industry

Table	Description	Name
ebt_inpr_msa	Index of Industrial production: Total Industry	c_d_e
ebt_inpr_msa	Index of Industrial production: Manufacturing	d
ebt_inpr_msa	Index of Industrial production: Manufacturing ind. working on orders	manuf_ord
ebt_inpr_msa	Index of Industrial production: Capital goods	capital
ebt_inpr_msa	Index of Industrial production: Intermediate goods	inter
ebt_inpr_msa	Index of Industrial production: Consumer durables	cons_dur
ebt_inpr_msa	Index of Industrial production: Consumer non-durables	cons_nondur
ebt_inpr_msa	Index of Industrial production: Consumer	consumer
ebt_inpr_msa	Index of Industrial production: Energy	energy
ebt_inpr_msa	Index of Industrial production: Mining and quarrying	c
ebt_inpr_msa	Index of Industrial production: Manufacture of food products	da
ebt_inpr_msa	Index of Industrial production: Manufacture of textiles	db
ebt_inpr_msa	Index of Industrial production: Manufacture of leather	dc
ebt_inpr_msa	Index of Industrial production: Manufacture of wood	dd
ebt_inpr_msa	Index of Industrial production: Manufacture of pulp, paper, etc.	de
ebt_inpr_msa	Index of Industrial production: Manufacture of coke, etc.	df
ebt_inpr_msa	Index of Industrial production: Manufacture of chemicals, etc	dg
ebt_inpr_msa	Index of Industrial production: Manufacture of rubber and plastic products	dh
ebt_inpr_msa	Index of Industrial production: Manufacture of other non-metallic	di
ebt_inpr_msa	Index of Industrial production: Manufacture of basic metals etc.	dj
ebt_inpr_msa	Index of Industrial production: Manufacture of machinery and equipment n.e.c.	dk
ebt_inpr_msa	Index of Industrial production: Manufacture of electrical and optical equipment	dl
ebt_inpr_msa	Index of Industrial production: Manufacture of transport equipment	dm
ebt_inpr_msa	Index of Industrial production: Manufacturing n.e.c.	dn
ebt_inpr_msa	Index of Industrial production: Electricity, gas and water supply	e
ebt_intv_m	Industry - Monthly turnover index: Manufacturing	d
ebt_intv_m	Industry - Monthly turnover index: Manufacturing ind. working on orders	manuf_ord
ebt_intv_m	Industry - Monthly turnover index: Capital goods	capital
ebt_intv_m	Industry - Monthly turnover index: Intermediate goods	inter
ebt_intv_m	Industry - Monthly turnover index: Consumer durables	cons_dur
ebt_intv_m	Industry - Monthly turnover index: Consumer non-durables	cons_nondur
ebt_intv_m	Industry - Monthly turnover index: Energy without section Nace E	energy_not_e
ebt_inno_m	Industry - Monthly indices of new orders - total: Manufacturing ind. w. on orders	manuf_ord
ebt_inno_m	Industry - Monthly indices of new orders - total: Consumer durables	cons_dur
ebt_inno_m	Industry - Monthly indices of new orders - total: Manufacture of leather etc.	dc
ebt_inno_m	Industry - Monthly indices of new orders - total: Manufacture of basic metals etc.	dj
ebt_inno_m	Industry - Monthly indices of new orders - total: Manufacture of machinery and eq. n.e.c.	dk
ebt_inno_m	Industry - Monthly indices of new orders - total: Manufacture of elec. and optical equipment	dl
ebt_inlb_howk	Industry - Volume of work done (hours worked): Total Industry	c_d_e
ebt_inlb_howk	Industry - Volume of work done (hours worked): Manufacturing	d
ebt_inlb_howk	Industry - Volume of work done (hours worked): Capital goods	capital
ebt_inlb_howk	Industry - Volume of work done (hours worked): Intermediate goods	inter
ebt_inlb_howk	Industry - Volume of work done (hours worked): Manufacture of rubber and plastic products	dh
ebt_inlb_howk	Industry - Volume of work done (hours worked): Manufacture of machinery and equip. n.e.c.	dk
ebt_inlb_howk	Industry - Volume of work done (hours worked): Manufacture of elec. and optical equipment	dl
ebt_inlb_howk	Industry - Volume of work done (hours worked): Manufacture of transport equipment	dm
ebt_inlb_wage	Industry - Gross wages and salaries: Total Industry	c_d_e
ebt_inlb_wage	Industry - Gross wages and salaries: Capital goods	capital
ebt_inlb_wage	Industry - Gross wages and salaries: Intermediate goods	inter
ebt_inlb_wage	Industry - Gross wages and salaries: Consumer non-durables	cons_nondur
ebt_inlb_wage	Industry - Gross wages and salaries: Consumer	consumer
ebt_inlb_wage	Industry - Gross wages and salaries: Energy	energy
ebt_inlb_wage	Industry - Gross wages and salaries: Mining and quarrying	c

Table 3: Monthly time series: Construction, Retail Trade

Construction		
Table	Description	Name
ebt_copr_m	Construction - Monthly production index (2000=100): Building and civil engineering	b4500
ebt_copr_m	Construction - Monthly production index (2000=100): Buildings	b4600
ebt_copr_m	Construction - Monthly production index (2000=100): Civil engineering	b4700
ebt_colb_m	Construction - Monthly indices of labour input (2000=100): Employment (n. of persons empl.)	empl
ebt_colb_m	Construction - Monthly indices of labour input (2000=100): Volume of work done (hours worked)	howk
ebt_colb_m	Construction - Monthly indices of labour input (2000=100): Gross wages and salaries	wage
ebt_cobp_m	Building permits - monthly data (2000=100) : Residential buildings	b4610

Retail Trade		
Table	Description	Name
ebt_ts_ret	Retail Trade. Index of Turnover: Retail trade, exc. motor veh., motorcycles; repair	g52
ebt_ts_ret	Retail Trade. Index of Turnover: Retail trade, except of motor vehicles, motorcycles	g52_not_g527
ebt_ts_ret	Retail Trade. Index of Turnover: Retail sale in non-specialized stores	g521
ebt_ts_ret	Retail Trade. Index of Turnover: Retail sale of food beverages or tobacco	g5211_g522
ebt_ts_ret	Retail Trade. Index of Turnover: Retail sale in non-specialized stores etc.	g5211
ebt_ts_ret	Retail Trade. Index of Turnover: Other retail sale in non-specialized stores	g5212
ebt_ts_ret	Retail Trade. Index of Turnover: Retail sale of non food products	g5212_g523_to_g526
ebt_ts_ret	Retail Trade. Index of Turnover: Retail sale of pharmaceutical, medical goods, cosmetic	g523
ebt_ts_ret	Retail Trade. Index of Turnover: Other retail sale of new goods in specialized stores	g524
ebt_ts_ret	Retail Trade. Index of Turnover: Retail sale of textiles, clothing, etc.	g5241_to_g5243
ebt_ts_ret	Retail Trade. Index of Turnover: Retail sale of household equipment	g5244_to_g5246
ebt_ts_ret	Retail Trade. Index of Turnover: Retail of books, newspapers and other	g5247_g5248
ebt_ts_ret	Retail Trade. Index of Turnover: Retail sale via mail order houses	g5261
ebt_ts_ret	Retail Trade. Index of Deflated Turnover: Retail trade, exc. motor veh., motorcycles; repair etc.	g52
ebt_ts_ret	Retail Trade. Index of Deflated Turnover: Retail trade, except of motor vehicles, motorcycles	g52_not_g527
ebt_ts_ret	Retail Trade. Index of Deflated Turnover: Retail sale in non-specialized stores	g521
ebt_ts_ret	Retail Trade. Index of Deflated Turnover: Retail sale of food beverages or tobacco	g5211_g522
ebt_ts_ret	Retail Trade. Index of Deflated Turnover: Retail sale in non-specialized stores etc.	g5211
ebt_ts_ret	Retail Trade. Index of Deflated Turnover: Other retail sale in non-specialized stores	g5212
ebt_ts_ret	Retail Trade. Index of Deflated Turnover: Retail sale of non food products	g5212_g523_to_g526
ebt_ts_ret	Retail Trade. Index of Deflated Turnover: Retail sale of pharmaceutical, medical goods, cosmetic	g523
ebt_ts_ret	Retail Trade. Index of Deflated Turnover: Other retail sale of new goods in specialized stores	g524
ebt_ts_ret	Retail Trade. Index of Deflated Turnover: Retail sale of textiles, clothing, etc.	g5241_to_g5243
ebt_ts_ret	Retail Trade. Index of Deflated Turnover: Retail sale of household equipment	g5244_to_g5246
ebt_ts_ret	Retail Trade. Index of Deflated Turnover: Retail of books, newspapers and other	g5247_g5248
ebt_ts_ret	Retail Trade. Index of Deflated Turnover: Retail sale via mail order houses	g5261
ebt_ts_ret	Retail Trade. Employment: Retail trade, exc. motor veh., motorcycles; repair etc.	g52
ebt_ts_careg	Car registrations (first registrations of private and commercial cars): Sale of motor vehicles	ea12

Table 4: Monthly time series: Monetary and Financial indicators, Labour market, External Trade

Table	Description	Name
Monetary and Financial indicators		
Table	Description	Name
mfrt_m	Exchange rates against the ECU/euro (average): Pound Sterling	mf-exa-rt_gbp
mfrt_m	Exchange rates against the ECU/euro (average): Yen (Japan)	mf-exa-rt_jpy
mfrt_m	Exchange rates against the ECU/euro (average): United States Dollar	mf-exa-rt_usd
mfrt_m	Nominal effective exchange rates (average): Pound Sterling	mf-exa-i_gbp
mfrt_m	Nominal effective exchange rates (average): Yen (Japan)	mf-exa-i_jpy
mfrt_m	Nominal effective exchange rates (average): United States Dollar	mf-exa-i_usd
mfma_m	Money supply M1 - SA	mf-m1-sa
mfma_m	Money supply M2 - SA	mf-m2-sa
mfma_m	Money supply M3 - SA	mf-m3-sa
mffa_m	Share price indices (average): Euro area (EA11-2000, EA12-2006, EA13)	mf-sp-i
irt_st_m	Money market interest rates - Monthly data: Day-to-day rates	mat_on
irt_st_m	Money market interest rates - Monthly data: 3-month rates	mat_m03
mfir_m	Interest rates - monthly data: Long term government bond yields - Maastricht definition (average)	mf-ltgby-rt
Labour market		
Table	Description	Name
une_rt_m	Harmonized unemployment rates, -/+ 25 years, monthly data: Total	total
une_rt_m	Harmonized unemployment rates, -/+ 25 years, monthly data: Less than 25 years	y0_24
une_rt_m	Harmonized unemployment rates, -/+ 25 years, monthly data: 25 years and over	y25_max
une_rt_m	Harmonized unemployment rates, -/+ 25 years, monthly data: Total Males	total
une_rt_m	Harmonized unemployment rates, -/+ 25 years, monthly data: Total Females	total
External Trade		
Table	Description	Name
etea12_m	External trade - Total Imports - Trade value	mio-eur-sa
etea12_m	External trade - Total Imports - Volume index (2000=100)	ivol-sa
etea12_m	External trade - Total Exports - Trade value	mio-eur-sa
etea12_m	External trade - Total Exports - Volume index (2000=100)	ivol-sa

Table 5: Monthly time series: EA 13 Business and consumer surveys (Source: DG ECFIN)

Table	Description	Name
bsin_m	Industry - monthly data: Production trend observed in recent months	bs-ipt
bsin_m	Industry - monthly data: Employment expectations for the months ahead	bs-ieme
bsin_m	Industry - monthly data: Assessment of order-book levels	bs-iob
bsin_m	Industry - monthly data: Production expectations for the months ahead	bs-ipe
bsin_m	Industry - monthly data: Industrial confidence indicator	bs-ici
bsbu_m	Construction - monthly data: Trend of activity compared with preceding months	bs-cta-bal
bsbu_m	Construction - monthly data: Assessment of order books	bs-cob-bal
bsbu_m	Construction - monthly data: Employment expectations for the months ahead	bs-ceme-bal
bsbu_m	Construction - monthly data: Price expectations for the months ahead	bs-cpe-bal
bsbu_m	Construction - monthly data: Construction confidence indicator	bs-cci-bal
bsrt_m	Retail sale - monthly data: Present business situation	bs-rpbs
bsrt_m	Retail sale - monthly data: Employment expectations for the months ahead	bs-ieme
bsrt_m	Retail sale - monthly data: Assessment of stocks	bs-ras
bsrt_m	Retail sale - monthly data: Expected business situation	bs-rebs
bsrt_m	Retail sale - monthly data: Employment	bs-rem
bsrt_m	Retail sale - monthly data: Retail confidence indicator	bs-rci
bsco_m	Consumers - monthly data: General economic situation over the last 12 months	bs-ges-ly
bsco_m	Consumers - monthly data: Price trends over the next 12 months	bs-pt-ny
bsco_m	Consumers - monthly data: Unemployment expectations over the next 12 months	bs-ue-ny
bsco_m	Consumers - monthly data: Savings at present	bs-sv-pr
bsco_m	Consumers - monthly data: Statement on financial situation of household	bs-sfsh
bsco_m	Consumers - monthly data: Consumer confidence indicator	bs-csmci