DOCUMENT

DE TRAVAIL

N° 232

ARE DISAGGREGATE DATA USEFUL FOR FACTOR ANALYSIS IN FORECASTING FRENCH GDP ?

Karim Barhoumi, Olivier Darné and Laurent Ferrara

January 2009

BANQUE DE FRANCE

EUROSYSTÈME

DIRECTION GÉNÉRALE DES ÉTUDES ET DES RELATIONS INTERNATIONALES

DIRECTION GÉNÉRALE DES ÉTUDES ET DES RELATIONS INTERNATIONALES

ARE DISAGGREGATE DATA USEFUL FOR FACTOR ANALYSIS IN FORECASTING FRENCH GDP ?

Karim Barhoumi, Olivier Darné and Laurent Ferrara

January 2009

Are disaggregate data useful for factor analysis in forecasting French GDP?*

Karim Barhoumi,† Olivier Darné‡ and Laurent Ferrara§

*We would like to thank A. Rua for helpful comments, as well as all the participants of the International Workshop on *Forecasting Macroeconomic Variables with Dynamic Factor Models* organized at the Banque de France, October 2008, and the participants of the *28th International Symposium on Forecasting*, Nice, June 2008. We would like also to thank V. Brunhes-Lesage for valuable research assistance. The views expressed herein are those of the authors and do not necessarily reflect those of the Banque de France.

†Banque de France, DGEI-DCPM, karim.barhoumi@banque-france.fr

‡Banque de France, DGEI-DCPM, and LEMNA, Université de Nantes, olivier.darne@univ-nantes.fr

§Banque de France, DGEI-DCPM, laurent.ferrara@banque-france.fr.

Abstract

This paper compares the GDP forecasting performance of alternative factor models based on monthly time series for the French economy. These models are based on static and dynamic principal components. The dynamic principal components are obtained using time and frequency domain methods. The forecasting accuracy is evaluated in two ways for GDP growth. First, we question whether it is more appropriate to use aggregate or disaggregate data (with three disaggregating levels) to extract the factors. Second, we focus on the determination of the number of factors obtained either from various criteria or from a fixed choice.

Keywords: GDP forecasting; Factor models; Data aggregation.
JEL Classification: C13; C52; C53; F47.

Résumé

Cet article compare les performances en prévision du PIB de différents modèles à facteurs dynamiques appliqués à un ensemble de données mensuelles représentatives de l'économie française. Les composantes principales dynamiques sont obtenues à partir de modèles estimés dans les domaines temporel et spectral. Les résultats en prévision du taux de croissance du PIB sont évalués sous deux angles différents. Dans un premier temps, nous déterminons empiriquement s'il est plus approprié d'utiliser des données agrégées ou désagrégées pour extraire les facteurs communs (nous considérons trois niveaux de désagrégation). Dans un second temps, nous nous intéressons à l'impact sur la prévision du choix du nombre de facteurs, soit en utilisant des critères statistiques, soit en fixant ce nombre de manière ad-hoc.

Mots-clés : Prévision du PIB; Modèles à facteurs dynamiques; Aggrégation.
Codes JEL : C13; C52; C53; F47.

Non-technical summary

Policy-makers and analysts are continually assessing the state of the economy. However, the most comprehensive measure of economic activity, namely GDP, is only available on quarterly basis with a delay of around 45 days, and often with significant revisions. In this respect, governments and central banks need to have an accurate and timely assessment of GDP growth rate for the current and the next quarters in order to provide a better and earlier analysis of the economic situation.

Recent works in the econometric literature consider the problem of summarizing efficiently a large set of variables (financial, hard and soft data, aggregated and disaggregated, ...) and using this summary for a variety of purposes including forecasting. Works in this field have been carried out in a series of recent papers by Stock and Watson (1999, 2002a, 2002b), Forni, Lippi, Hallin and Reichlin (2000, 2001, 2004, 2005), Doz, Giannone and Reichlin (2006, 2007) or Giannone, Reichlin and Small (2008). Factor analysis has been the main tool used in summarizing the large datasets. Under the factor model approach each time series is represented as the sum of two orthogonal components: the common component, which is strongly correlated with the rest of the panel and is a linear combination of the factors, and the idiosyncratic component. The common component of the time series is driven by a few underlying uncorrelated and unobservable common factors.

In this paper, we compare the GDP forecasting performance of alternative factor models based on monthly time series for the French economy. These models are based on static and dynamic principal components. The dynamic principal components are obtained using time and frequency domain methods. The forecasting accuracy is evaluated in two ways for the GDP growth. First, we question whether it is more appropriate to use aggregate or disaggregate data (with three disaggregating levels) to extract the factors. Second, we focus on the determination of the number of factors obtained either from various criteria or from a fixed choice.

From this application on the French GDP growth rate, we can conclude that complex dynamic models with strongly disaggregated data base do not necessarily lead to the best forecasting results. Indeed, the simple static Stock and Watson (2002a) approach with an aggregated data base of 20 series lead to comparable forecasting results when using a disaggregated data base of 140 series with a dynamic model. Moreover, we empirically show that the use of Bai and Ng (2002, 2007) tests would lead to unefficient forecasting results and that the inclusion of a higher number of factors improves the performances.

Résumé non-technique

Les décideurs politiques et les analystes économiques et financiers cherchent à évaluer de manière continue les fluctuations de l'économie. Toutefois, la mesure la plus complète de l'activité économique, à savoir le PIB, n'est disponible que sur une fréquence trimestrielle et avec environ 45 jours de délai. Par conséquent, les gouvernements et les banques centrales ont besoin d'avoir à leur disposition une évaluation rapide et fiable du taux de croissance du PIB, pour le trimestre en cours et les trimestres suivants.

Des travaux récents de la littérature économétrique ont considéré le problème de la réduction de la dimension d'un grand ensemble de données (enquêtes d'opinion, activité économique, données financières, données agrégées et désagrégées par secteur ou par pays ...) et de l'utilisation de ces variables synthétiques pour différents objectifs, en particulier la prévision macroéconomique. Des travaux de recherche sur ce champ d'application ont été menés par Stock et Watson (1999, 2002a, 2002b), Forni, Lippi, Hallin et Reichlin (2000, 2001, 2004, 2005), Doz, Giannone et Reichlin (2006, 2007) ou Giannone, Reichlin et Small (2008). L'analyse factorielle est le principal outil utilisé dans ces travaux pour résumer un grand ensemble de données. Cette méthode considère que chaque variable peut être représentée comme la somme de deux composantes orthogonales: une composante commune, combinaison linéaire des variables et fortement corrélée avec le reste des variables, et une composante idiosyncratique.

Dans ce document, nous comparons les performances en prévision du PIB français de différents modèles à facteurs, statiques et dynamiques, appliqués à des données mensuelles. La précision de la prévision est évaluée autour de deux axes. D'abord, nous nous demandons s'il est plus approprié d'utiliser des données agrégées ou des données désagrégées (avec trois niveaux de désagrégation). Ensuite, nous nous intéressons au choix du nombre de facteurs obtenu soit à l'aide de critères statistiques, soit par détermination *a priori*.

A partir des résultats obtenus sur le taux de croissance du PIB français, nous concluons qu'un modèle à facteurs intégrant une dynamique complexe, ajusté à une grande base fortement désagrégée, ne fournit pas nécessairement les meilleures prévisions. En effet, l'approche simple de Stock et Watson (2002a) associée à une base de données de 20 variables conduit à des résultats similaires à un modèle à facteurs dynamique appliqué à 140 variables. De plus, nous montrons de manière empirique que les tests de Bai et Ng (2002, 2007) peuvent mener à des résultats inefficaces en prévision et que l'inclusion d'un nombre plus élevé de facteurs améliore les perfomances.

1 Introduction

Policy-makers and analysts are continually assessing the state of the economy. However, gross domestic product [GDP] is only available on quarterly basis with a delay of 1.5 months (45 days), and often with significant revisions. In this respect, governments and central banks need to have an accurate and timely assessment of GDP growth rate for the current and the next quarters in order to provide a better and earlier analysis of the economic situation.

Economists and forecasters nowadays typically have access to information scattered through huge numbers of observed time series – hard and soft, aggregated and disaggregated, real and nominal variables.

Recent works in the econometric literature consider the problem of summarizing efficiently a large set of variables and using this summary for a variety of purposes including forecasting. Works in this field have been carried out in a series of recent papers by Stock and Watson (1999, 2002a, 2002b), Forni, Lippi, Hallin and Reichlin (2000, 2001, 2004, 2005), Doz, Giannone and Reichlin (2006, 2007) or Giannone, Reichlin and Small (2008). Factor analysis has been the main tool used in summarizing the large datasets. Under the factor model approach each time series is represented as the sum of two orthogonal components: the common component, which is strongly correlated with the rest of the panel and is a linear combination of the factors, and the idiosyncratic component. The common component of the time series is driven by a few underlying uncorrelated and unobservable common factors. In the classic or exact factor model, idiosyncratic components are mutually uncorrelated (orthogonal idiosyncratic elements), limiting thus economic applications.

In traditional factor analysis, for a given size of the cross-section n (i.e. small n), the model can be consistently estimated by maximum likelihood. The literature has proposed both frequency domain (Geweke, 1977; Sargent and Sims, 1977; Geweke and Singleton, 1980) and time domain (Engle and Watson, 1981; Stock and Watson, 1989; Quah and Sargent, 1992) methods. It is assumed that there is no cross-correlation among the idiosyncratic components at any lead and lag. This assumptions allows for identification of common and idiosyncratic components but represents a strong restriction.

Recent advances in the theory of dynamic factor model [DFM] have generalized the idea of factor analysis to handle less strict assumptions on the covariance of the idiosyncratic elements (approximate factor structure) and proposed non-parametric estimators of the common factors based on principal components, which is feasible for n

large. They have shown that, under suitable technical conditions, it is possible to estimate the dynamic factors consistently in an approximate dynamic factor model when the time series (T) and cross-sectional (n) dimensions are large (Forni et al., 2000; Stock and Watson, 2002a, 2002b). The extensions of DFM to large n can therefore be viewed as a particularly efficient way of extracting information from a large number of data series. Furthermore, these models differ from the classic factor model in that they allow the idiosyncratic errors to be weakly serial and cross-sectional correlated to some extent.

In their seminal papers, Stock and Watson (1999, 2002a, 2002b) [SW] show that, if the data can be described by an approximate dynamic factor model, then under certain conditions (restrictions on moments and nonstationarity) the latent factors can be estimated consistently by the principal components of the sample covariance matrix. Stock and Watson (2002a) also provide conditions under which these estimated factors can be used to construct asymptotically efficient forecasts by a second stage forecasting regression in which the estimated factors are the predictors. Otherwise, their forecast is based on a projection onto the space spanned by the static principal components of the data. Thus, being based on an eigenvalue decomposition of the contemporaneous covariance matrix only, their approach does not exploit the dynamic relations between the variables of the panel.

To take into account a richer dynamic structure for the factor models, various extensions [1] to the static principal component estimators have been developed either in the time domain or in the frequency domain [2].

Doz et al. (2006, 2007) [DGR] propose the implementation of the common factors as unobserved components in a state-space form. Factor dynamics is therefore modelled explicitly. In Doz et al. (2007) they introduce a parametric time domain two-step estimator involving principal components and Kalman filter to exploit both factor dynamics and idiosyncratic heteroscedacticity. In the first step, the parameters of a dynamic approximate factor are first estimated using a simple least squares on principal components. In the second step, the factors are estimated via the Kalman smoother.

[1] See Reichlin (2003), Stock and Watson (2006), Breitung and Eickmeier (2006), Eickmeier and Ziegler (2008) for a survey on factor models. Kapetanios and Marcellino (2004) and Schumacher (2007) compare factor estimation techniques.

[2] Another dynamic factor model approach have been proposed by Kapetanios (2004), Camba-Mendez and Kapetanios (2005) and Kapetanios and Marcellino (2004), based on subspace algorithms for state-space models, but it is not considered in this study. See Schumacher (2007) and Eickmeier and Ziegler (2008) for a comparison of this approach with others dynamic factor models.

This procedure allows to consider dynamics in the factors and heteroskedasticity in the idiosyncratic variance. In Doz et al. (2006) they suggest a quasi maximum likelihood estimation [QML], in the sense of White (1980), for the approximate factor model in large panels. They show that traditional factor analysis in large cross-section n is feasible and that consistency is achieved even if the underlying data generating process is an approximate factor model rather than an exact one. The misspecification error due to the approximate structure of the idiosyncratic component vanishes asymptotically for n and T large, provided that the cross-correlation of the idiosyncratic processes is limited and that the common components are pervasive throughout the cross section as n increases.
Forni et al. (2000, 2001, 2004, 2005) [FHLR] use dynamic principal component analysis in the frequency domain to estimate large-scale factor models, where they estimate the common factors based on generalized principal components in which observations are weighted according to their signal-to-noise ratio. This model is also called generalized dynamic factor model [GDFM]. FHLR dynamic principal components are based on the spectral density matrix (i.e. dynamic covariations in the frequency domain) of the data and consequently are averages of the data weighted and shifted through time. This method incorporates an explicitly dynamic element in the construction of the factors.

In the recent applied macro-economics literature, especially the macro-economic forecasting literature, factor models with large dataset have received increasing attention [3]. Literature has not yet reached a consensus between static and dynamic principal component approaches. Using a large panel of US macroeconomic variables, Stock and Watson (2006) and D'Agostino and Giannone (2007) find that SW and FHLR methods perform similarly, while Boivin and Ng (2005) find that SW's method largely outperforms the FHLR's and, in particular, they conjecture that the dynamic restrictions implied by the latter method are harmful for the forecast accuracy of the model. Schumacher (2007) finds mixed results between the SW and FHLR's methods in forecasting German macroeconomic variables. However, there little empirical comparison between the SW, FHLR and DGR methods in forecasting, except Barhoumi et al. (2008).

[3] Moreover, various applications using DFM provided additional favorable evidence for the forecasting accuracy of the factors models (e.g., Brisson et al., 2003; Camacho and Sancho, 2003; Artis et al., 2005; Cheung and Demers, 2007).

A feature stressed in recent applications of factor models is the use of data from large panels. Because the theory is developed for large n and T, there is a natural tendency for researchers to use as much data as are available. However, some studies suggests that n does not need to be extremely large for the principal components estimator to give reasonably precise estimates (Watson, 2003; Bai and Ng, 2002; Boivin and Ng, 2006) [4]. Therefore, from a forecasting point of view, we question whether it is more appropriate to use aggregate or disaggregate data, with two disaggregating levels, to extract the factors from various DFMs.
As suggested by Schumacher (2007), performance-based model selection as well as information criteria are used for model specification. For the model selection using information criteria, we use criteria by Bai and Ng (2002) and Bai and Ng (2007) for the number of static and dynamic factors, respectively. We also consider a *a priori* fixed choice of the number of factors, by increasing progressively this number, to forecast GDP. The forecasting accuracy of alternative factor models introduced above is discussed in this paper.

2 Factor models

2.1 The strict factor model

In the factor model framework, variables are represented as the sum of mutually orthogonal unobservable components: the common component and the idiosyncratic component. The common component is driven by a small number of factors common to all the variables in the model. The idiosyncratic component is driven by variable-specific shocks. In an r-factor model each element of a vector $X_t = [x_{1t}, ..., x_{nt}]^{'}$ can be represented as:

$$\begin{aligned} x_{it} &= \lambda_{i1}F_{1t} + ... + \lambda_{ir}F_{rt} + \xi_{it}, \;\; t = 1, ..., T \\ &= \lambda_i^{'}F_t + \xi_{it} \end{aligned} \tag{1}$$

where $\lambda_i = [\lambda_{i1}, ..., \lambda_{ir}]^{'}$, F_t is a vector of r common factors such that $F_t = [F_{1t}, ...F_{rt}]^{'}$ and $\xi_t = [\xi_{1t}, ..., \xi_{nt}]^{'}$ is a vector of n idiosyncratic mutually uncorrelated components. More compactly, the model (1) can be rewritten as:

[4] Watson (2003) found that the marginal gain (in terms of forecast mean-squared error) from increasing n beyond 50 appears less substantial. Bai and Ng (2002) found that in simulations, the number of factors can be quite precisely estimated with n as small as 40 when the errors are iid. Boivin and Ng (2006) showed that, in simulations and the empirical examples, the factors extracted from as few as 40 series seem to do no worse, and in many cases, better than the ones extracted from 147 series.

$$X_t = \Lambda F_t + \xi_t, \quad (2)$$

where Λ is the loading matrix such that $\Lambda = [\lambda_1, ..., \lambda_n]'$.
In the framework of a strict factor model, it is also assumed that ξ_t is a serially uncorrelated vector such that $E(\xi_t) = 0$ and for any given i, $E(\xi_{it}\xi_{it'}) = 0$ if $t \neq t'$ and $E(\xi_{it}\xi_{it'}) = \sigma_i^2$ otherwise. In addition, it is assumed that $E(F_t) = 0$ and $E(F_t F_t^{'}) = \Omega$ and that the factors are uncorrelated with the idiosynchractic noise. From these assumptions it follows that:

$$E(X_t X_t^{'}) = \Lambda\Omega\Lambda^{'} \quad (3)$$

It can be shown that the least-squares estimate of the loading matrix Λ is also the principal components (PC) estimate.

In traditional factor analysis, for a small size of the cross-section n, the model can be consistently estimated by maximum likelihood. The literature has proposed both frequency domain and time domain methods. In frequency domain, Sargent and Sims (1977) and Geweke (1977) were the first to propose a dynamic factor model. They obtain parameter estimates by maximizing the spectral likelihood function. In time domain, Engle and Watson (1981) propose the use of Fisher scoring to maximize the likelihood in the time domain and apply this method to a one-factor model. Watson and Engle (1983) and Quah and Sargent (1993) adopt the expectation-maximization (EM) algorithm of Dempster et al. (1977) to estimate a factor model [5].

2.2 Approximate factor model

The fairly restrictive assumption of the strict factor model can be relaxed if it is assumed that the number n of variables tends to infinity (Chamberlain and Rothshield, 1983; Connor and Korajczyk, 1986, 1988, 1993; Stock and Watson, 2002a; Bai 2003). First, it is possible to allow for (weak) serial correlation of the idiosyncratic errors. Thus, the principal component estimator remains consistent if the idiosyncratic errors are generated by stationary short-memory ARMA processes. However, persistent and non-ergodic processes, such as the random walk, are ruled out. Second, the idiosyncratic errors may be weakly cross-correlated and heteroskedastic. This allows for finite "clusters of correlation" among the errors. Another way to express this assumption is

[5] The EM algorithm has the advantage that it is stable and it is sure to converge to an optimum. However, Watson and Engle (1983) found that convergence is often slow.

to assume that all eigenvalues of $E(\xi_t \xi_t') = \Sigma$ are bounded. Third, the model allows for weak correlation between the factors and the idiosyncratic components.

2.2.1 Stock and Watson (2002)

In order to derive the factor, Stock and Watson (2002a, 2002b) [SW] use static principal component analysis. The aim of the static component analysis is to choose the parameters and factor for the model (2) in order to maximize the explained variance of the original variables for a given small r of factors F_t. Under some technical assumptions (restrictions on moments and nonstationarity), the column space spanned by the dynamic factors F_t can be estimated consistently by the (static) principal components of the covariance matrix of the X's. The principal component estimator is computationally convenient, even for very large n. More precisely, we consider $\widehat{\Gamma}_0 = (1/T)\sum_{t=1}^{T} X_t X_t'$ as an estimation of the contemporaneous variance-covariance matrix of the vector of the time series X_t. The aim of this approach is to find r linear combinations of the time series data $\widehat{F_{j,t}} = \widehat{S}_j' X_t$ for $j = 1,...,r$ that maximize the variance of the factors $\widehat{S}_j' \widehat{\Gamma}_0 \widehat{S}_j$. Due to the fact that the number of the factors should be sufficiently small compared with the total number of time series, $r \ll n$, SW impose the normalization $\widehat{S}_j' \widehat{S}_j = 1$ for $i = j$ and 0 for $i \neq j$.

Hence, the maximization problem can converted to an eigenvalue problem:

$$\widehat{\Gamma}_0 \widehat{S}_j = \widehat{\mu}_j \widehat{S}_j, \tag{4}$$

where $\widehat{\mu}_j$ denotes the j-th eigenvalue and $\widehat{S}_j$ its $(N \times 1)$ corresponding eigenvector. As before, after the calculation of the maximum of n eigenvalues, they are ranked in decreasing order of magnitude and the eigenvectors according to the r largest eigenvalues are weights on the static factors:

$$F_t^{SW} = \widehat{S}' X_t, \tag{5}$$

where $\widehat{S}$ is the $(n \times r)$ matrix of stacked eigenvectors $\widehat{S} = (\widehat{S}_1, ..., \widehat{S}_r)$. We need only one auxiliary parameter r to derive the factors.

2.2.2 Dynamic Factor Models

To integrate dynamics in forecasting, SW propose to apply an autoregressive model to the factors. Another way to proceed to take dynamics into account is to model explicitly the dynamics of the factors F_t. More precisely, we assume that the dynamic factor model representation is given by:

$$X_t = \chi_t + \xi_t, \tag{6}$$

where the component χ_t integrates a linear dynamic such that:

$$\chi_t = A(L)F_t, \tag{7}$$

where $A(L)$ is a $(n \times r)$ matrix describing the autoregressive form of the r factors. If we assume that there exists a $(n \times q)$ matrix $B(L)$ such that $B(L) = A(L)N(L)$ with $N(L)$ of dimension $(r \times q)$, then the dynamic factor is such that $F_t = N(L)U_t$ where U_t is a $(q \times 1)$ independent vector containing the dynamic shocks. From equations (6) and (7) it follows that the factor dynamics are described by:

$$A(L)F_t = B(L)U_t \tag{8}$$

Equation (8) specifies a VAR(p) model for the factor F_t with lag polynomial $A(L) = \sum_{i=1}^{p} A_i L^i$. F_t is thus the $(r \times 1)$ vector of the stacked factors with $r = q(p+1)$.

Doz et al. (2006, 2007)

In two successive papers, Doz *et al.* (2006, 2007) [DGR] proposed a dynamic factor model for a large set of data based on a state-space representation. More precisely, DGR propose two approaches to estimate the dynamic factor model: the two-steps approach (2007) and the QML approach (2006). We briefly present those estimation methods below.

The two-steps approach consists in first estimating the parameters by principal component. Then, in the second step, the factors are estimated via Kalman smoothing. DGR (2007) cast the model into a state-space form with equations (6)-(7) referring to the state equation and equation (8) referring to the space equation.

For a given number of factors r and dynamic shocks q, the estimation proceeds in the following two steps. In the first step, we estimate $\widehat{F_t}$ using principal component analysis as initial estimate. Then, we estimate $\widehat{\Lambda}$ by regressing X_t on the estimated factors $\widehat{F_t}$. The covariance matrix of the idiosyncratic components $\widehat{\xi_t} = X_t - \widehat{\Lambda}\widehat{F_t}$ denoted as $\widehat{\sum_\xi}$ is also estimated. The estimation of a VAR(p) model for the factors $\widehat{F_t}$ yields $\widehat{A}(L)$ and the residual covariance of $\widehat{\varsigma_t} = \widehat{A}(L)\widehat{F_t}$ denoted $\widehat{\sum_\varsigma}$. To obtain an estimate of $N(L)$, given the number of dynamic shocks q, DGR (2007) apply an eigenvalue decomposition of $\widehat{\sum_\varsigma}$. Let M be the $(r \times q)$-dimensional matrix of the eigenvectors corresponding to the q largest eigenvalues and let the $(q \times q)$-dimensional matrix P contain the largest eigenvalues on the main diagonal and zero otherwise. Then the estimate of $N(L)$ is $\widehat{N}(L) = M \times P^{-1/2}$.

In the second step, the coefficients and auxiliary parameters of the system of equations (6), (7) and (8) are fully specified numerically. The model is cast into state-space form and the Kalman smoother yields new estimates of the factors.

DGR (2006) propose a second approach based on quasi-maximum likelihood [QML] estimations of a dynamic approximate factor model [6].The central idea is to treat the exact factor model as a misspecified approximating model and analyze the properties of the maximum likelihood estimator of the factors under misspecification, that is when the true probabilistic model is approximated by a more restricted model. This is a QML estimator in the sense of White (1980). Maximum likelihood is analyzed under different sources of misspecification such as omitted serial correlation of the observations, cross-sectional correlation of the observations and cross-sectional correlation of the idiosyncratic components. They show that the effects of misspecification on the estimation of the common factors is negligible for large sample size (T) and the cross-sectional dimension (n). The estimator is then a valid parametric alternative to principal components which can potentially produce efficiency improvements due to the exploitation of the factor dynamics and the non sphericity of the idiosyncratic components.

The model defined in equations (6), (7) and (8) can be cast into a state-space form with the number of states equal the number of common factors r. For any set of parameters the likelihood can then be evaluated using the Kalman filter. Given the QML estimates of the parameters θ of the model, the common factors can be approximated by their expected value, which can be computed using the Kalman smoother. [7]

Throughout this paper, we attribute the following notation F_t^{2S} to this first approach and F_t^{QML} for the second approach.

Forni et al. (2004, 2005)

To estimate the dynamic factors and their covariances, FHLR (2000, 2001, 2004, 2005) propose dynamic principal analysis in the frequency domain. The dynamic prin-

[6] Recently, Junbacker and Koopman (2008) propose new results for the likelihood-based analysis of the dynamic factor model. The estimation of the factors and parameter estimation is obtained by maximum likelihood and Bayesian methods using Markov chain Monte Carlo approach.

[7] The likelihood can be maximized via the EM algorithm which requires at each iteration only one run of the Kalman smoother.

cipal components are derived in order to maximize the common components' variance under orthogonality restrictions. The optimization leads to a dynamic eigenvalue problem of the spectral density matrix of the vector of observed variables. The spectral density matrix of the vector of observed variables $\Sigma(\theta)$ of X_t is estimated using the frequency domain representation of the time series. For each frequency θ lying on the interval $[-\pi, \pi[$, dynamic principal components are obtained through the dynamic eigenvector and eigenvalue decomposition of the spectral density matrix [8].

The common components are the orthogonal projections of the data on the present, past and future of the q dynamic principal components. The projection coefficients of the common components, $A(L)$, are the result of an inverse Fourier transform of the first q dynamic eigenvectors. More precisely, this transformation translates the results found in the spectral domain (dynamic eigenvectors) into a filter in the time domain $A(L)$. The frequency domain estimator yields a two-sided filter. Consequently, problems arise at the end of the sample since future observations are needed to estimate the common components. To solve this problem FHLR (2005) suggest a refinement of their procedure that retains the advantages of the dynamic approach, while the common component is based on a one-sided filter. Following this procedure, the factor space is approximated by r static aggregates instead of q dynamic principal components. These r contemporaneous averages are however based on the information of the dynamic approach.

The procedure consists in two steps. In the first step, it relies on the dynamic approach, which delivers estimates of the covariance matrices of the common and idiosyncratic component, $\hat{\Gamma}_\chi(\theta)$ and $\hat{\Gamma}_\xi(\theta)$, through an inverse Fourier transform of the spectral density matrices. The covariance of common components is obtained by

$$\hat{\Gamma}_{\chi,k}(\theta) = \frac{1}{2M+1} \sum_{h=0}^{2M} \hat{\Sigma}_\chi(\theta_h) e^{ik\theta_h} \tag{9}$$

for $k = -M, \ldots, M$. The covariance of idiosyncratic component can be obtained accordingly.

In the second step, this information is used to construct the factor space by r contemporaneous averages, wherein the variables are weighted according to their common/idiosyncratic variance ratio obtained from the contemporaneous covariance matrices estimated in the first step. These r aggregates are the solutions from a generalized principal component problem and have the efficient property of reducing the

[8]The eigenvalue-eigenvector decomposition also allows to split up the spectral density matrix into a spectral density matrix of the common component $\Sigma_\chi(\theta)$ and spectral density matrix of idiosyncratic component $\Sigma_\xi(\theta)$.

idiosyncratic disturbances in the common factor space to a minimum, by selecting the variables with the highest common/idiosyncratic variance ratio. The number of aggregates is equal to $r = q(p+1)$, which is the static rank of the spectral density matrix of the common factors, p indicates the order of the lag operator in $A(L) = \sum_{i=1}^{p} A_i L^i$ for equation (8).

FHLR (2005) stipulate that the maximization problem, in order to find the r aggregates, can be represented as a generalized eigenvalue problem

$$\hat{\Sigma}_{\chi,0}\hat{Z}_j = \hat{\mu}_j\hat{\Sigma}_{\xi,0}\hat{Z}_j \tag{10}$$

where $\hat{\mu}_j$ denotes the j-th generalized eigenvalue, $\hat{Z}_j$ its $(n \times 1)$ corresponding eigenvectors, and $\hat{\Sigma}_{\chi,0}$ and $\hat{\Sigma}_{\xi,0}$ are the contemporaneous variance-covariances of the dynamic and idiosyncratic components, respectively. Note that FHLR (2005) impose the following normalization $\hat{Z}_j'\hat{\Sigma}_{\xi,0}\hat{Z}_i = 1$ for $i = j$ and 0 for $i \neq j$. Last, the n eigenvalues are ranked by in decreasing order of magnitude, the factors are obtained as the product of r eigenvectors corresponding to the largest eigenvalues and the vector of observable variables X_t such as:

$$F_t^{FHLR} = \widehat{Z}'X_t \tag{11}$$

where $\widehat{Z} = (\widehat{Z_1}, ..., \widehat{Z_r})$ is the $(n \times r)$ matrix of the stacked eigenvectors.

3 Forecasting with factor model

In this section we compare the four previously factor estimation methods in order to forecast the French GDP growth rate one-quarter ahead, by using the same data base disaggregated for three various levels.

3.1 Forecast equation

In order to evaluate the predictive content conveyed by the factor estimates, they have to be implemented into a forecasting model. We use the four types of estimated factors previously presented, namely F_t^{SW}, F_t^{2S}, F_t^{QML} and F_t^{FHLR}, for prediction in a dynamic model. In this paper, we focus on the one-step-ahead prediction of the French GDP growth rate, denoted $\hat{Y}_{t+1}$. As for example in Forni *et al.* (2003a), Kapetanios and Camba-Mendez (2004) or Schumacher (2007), we estimate the one-step-ahead predictor by using the following leading equation:

$$\hat{Y}_{t+1} = \beta' F_t + \phi(L) Y_t, \tag{12}$$

where F_t is the r-vector of estimated factors obtained by using one of the four methods, $\beta = (\beta_1, \ldots, \beta_r)'$ is a coefficient vector of length r and $\phi(.)$ is a polynomial of order p. The $r + p + 1$ parameters of the model, namely $(\beta_1, \ldots, \beta_r, \phi_0, \phi_1, \ldots, \phi_p)$, are estimated by ordinary least-squares.

In order to compare with the factor-augmented approach (equation (12)), we consider two simple benchmark predictors. First, we use the naive predictor such that $\hat{Y}_{t+1} = Y_t$ and second the autoregressive predictor given by:

$$\hat{Y}_{t+1} = \psi(L) Y_t, \tag{13}$$

Significant lags up to the 4^{th} order with an associated probability of the t-stat of less than 5% were kept in the AR(p) polynomial $\psi(.)$.

3.2 Data description

As one of our aim is to assess the effects of data disaggregation on forecasting performances, we construct three different monthly data bases that we called small, medium and large, starting from the same set of data for the French economy. The small data base consists in 20 variables including hard data (manufacturing industrial production index, consumer spending, new cars registrations, selling of industrial vehicles, housing starts, imports and exports), soft data (industrial confidence index, consumer confidence index, services confidence index, retail sales, European Commission surveys on assessment of order-books levels for both domestic and foreign demand and production expectations for the months ahead), financial data (French stock index, long-term, short-term rate and housing interest rates) and prices (oil price and consumer price index). Surveys in the industry and services are provided by the Banque de France and the consumer survey stems from Insee, the French national statistical institute. From this small data base, we first decide to disaggregate soft data, when possible, according to their various questions, instead of using composite index as in the small data base. That is, we split the three confidence indicators (industry, services and consumers) according to the first-level questions. By doing this, we extend the base to 51 variables, denoted as the medium data base. Last, we decide to carry out a sectoral disaggregation of the data when possible. For example, we split the industrial sector into consumer goods, equipment goods, intermediate goods, agri-food goods and car industry. The large data base consists thus in 140 variables. When necessary, data have been differenced to avoid a non-stationary component. Last, data have been centered and standardized before entering the factor model.

For each data base (small, medium and large), we extract the r common factors by using the four extraction methods previously described. We fixed a priori $r = 5$ and we will compare the effects of the number of factors on forecasting GDP, by comparison with a pre-specified number of factors estimated with the Bai and Ng test (2002, 2007). Moreover, as the explained variable, GDP growth rate, is quarterly, we average the monthly estimated factors into quarterly factors in order to estimate the predicted value through equation (12).

3.3 Forecasting results

Out-of-sample rolling forecasts are carried out to determine the predictive power of each factor extraction method. The rolling forecasts have been implemented over the period 2000q1-2007q4. Parameters of the model are re-estimated at each step when new data are included in the learning set. Concerning the specification of the models, we keep the statistically significant models as regards the number of autoregressive lags and the number of factors involved in the leading equation (12), by using Student tests on parameter estimates with a confidence level of 95 %. Moreover, we check the robustness of the models by assuring that parameters are significant through time.

To assess the predictive accuracy, we use the classical root mean-squared error (RMSE, henceforth) criterion defined by the following equation:

$$RMSE(i) = \sqrt{\frac{1}{h}\sum_{t=1}^{h}(Y_{t+1} - \hat{Y}^{i}_{t+1})^2}, \qquad (14)$$

where $i \in \{SW, 2S, QML, FHLR\}$, h is the number of quarters considered in the rolling forecast exercise ($h = 32$ from Q1 2000 to Q4 2007), Y_t is the true value of the GDP growth rate. Note that we use as true values, the chain-linked values released in February 2008 by the quarterly national accounts of the French national statistical institute.

Results in terms of RMSE are presented in Table 1. This table also contains the optimal specification of the models (12), in the sense that all the presented models are statistically significant. When several models have been found significant, we retain the one providing with the lower RMSE. From those results, we conclude first that factor-augmented models clearly outperform naive and autoregressive models, indicating thus that the information conveyed by the factors is useful. Second, we observe

Base	Predictor	RMSE	AR lags	Factors
	Naive	0.5032		
	AR	0.4039	1,2	
Small	SW	0.2314	2,3	F_{1t} F_{2t}
	2S	0.2474	2	F_{1t} F_{2t}
	QML	0.2442	2	F_{1t} F_{2t}
	FHLR	0.2466	2,3	F_{2t} F_{4t} F_{5t}
Medium	SW	0.2382	1	F_{1t} $F_{1(t-1)}$ F_{3t} F_{4t} F_{5t}
	2S	0.2400	2	F_{1t} F_{4t}
	QML	0.2631	2	F_{1t} F_{4t}
	FHLR	0.2556	3	F_{1t} F_{5t}
Large	SW	0.2357	2	F_{1t} F_{2t} $F_{2(t-1)}$ F_{5t}
	2S	0.2391	3	F_{1t} F_{2t} F_{5t}
	QML	0.2642	3	F_{1t} F_{2t} F_{5t}
	FHLR	0.2758	3	F_{2t} F_{5t}

Table 1: RMSEs for the 3 data bases and the four factor extraction methods over the period Q1 2000 - Q4 2007

that the simplest method as regards parameter estimation, namely the SW approach, always provides the best results for a given data base, although the difference with the worst results is not huge (lower than 0.04 points). Modified Diebold-Mariano tests of Harvey, Leybourne and Newbold (1997) have been carried out in order to test the equality of forecast performances (see results in Table 2). With a confidence level of 90%, we cannot conclude that results from SW approach are statistically different from those of other approaches. This result appears interesting for practitioners in search for parsimony and simplicity in modelling when they are in charge of providing results on a regular and frequent basis. Third, we observe that the enlargement of the data base does not have a strong impact on forecasting accuracy. For example, the SW approach leads roughly to the same forecasting error, although the structure of the model is changing with the base. Indeed, for the medium and large bases, a dynamic is needed and higher factors are included. As regards the FHLR approach, the forecast accuracy decreases when the data base widens, which is a striking result. To a certain extent, the QML approach provides also with the same result. This results means that filling the factor model with the largest as possible data set is not necessarily the best strategy. This result is similar to those found in Bai and Ng (2002), Watson (2003) and Boivin and Ng (2006). A limited choice of data, along with the choice of the optimal model in terms of specification in equation (12), can lead to similar or even best results.

This latter remark leads us to question on the way to specify optimally the leading equation (12). Especially, the number r of factors to include is always questionable. In the econometrics literature, a classical answer to this issue consists in using the Bai and Ng (2002, 2007) tests, who suggest information criteria to estimate consistently the number of factors as n and T tend to infinity. The 2002 paper deals with static factor models while the 2007 paper concerns dynamic factor models. To compare the impact on forecasting accuracy of the choice of the number of factors in equation (12), we consider first the tests of Bai and Ng (2002, 2007), then we adopt a naive sequential approach which consists in using a rolling procedure among the factors and in comparing the resulting RMSEs. We use a sequential approach that integrates first only the first factor, then the two first factors, then the three first factors, etc. We do not exceed five factors. In this experience, we present all the estimated models even if the factors are not significant. Results are presented in Table 3. From this table, it is noteworthy that the number of factors has a strong influence on the forecasting accuracy. Indeed, it turns out that we cannot limit to the two or three first factors as generally invoked in the Bai-Ng tests. The application of those tests on our data lead to retain only two factors for the small and medium bases and three factors for the large base, both for static and dynamic approaches. Yet the inclusion of the 4th factor may allow a strong reduction of the RMSE, for example for the FHLR method with the small data base (0.3491 against 0.2701) or for the SW method with the medium data base (0.3628 against 0.2757). Moreover, the 5th factor may also have a strong influence on prediction as it is the case for all the methods when using the large data base (e.g., 0.2783 for the 2S method against 0.2577 for the SW method). Therefore, for the large data base, it seems that high orders factors may contain a predictive power and not include them in forecasting, as it is the case when using the Bai-Ng tests, may lead to inaccurate results.

Base	2S	QML	FHLR
Small	0.2249	0.2627	0.1108
Medium	0.4675	0.1530	0.3087
Large	0.4343	0.1950	0.1295

Table 2: P-values of Modified Diebold-Mariano tests (Harvey, Leybourne and Newbold, 1997) against the SW model, over the period Q1 2000 - Q4 2007 ($h = 32$ observations). If the P-value is lower than the type I risk α equal to, for example, 0.05, it means that we can reject the null hypothesis of equality of expected forecast performance with a risk α.

Base	Method	F_{1t}	F_{1t}, F_{2t}	F_{1t}, F_{2t}, F_{3t}	$F_{1t}, F_{2t}, F_{3t}, F_{4t}$	$F_{1t}, F_{2t}, F_{3t}, F_{4t}, F_{5t}$
Small	SW	0.2473	0.2326	0.2332	0.2376	0.2361
	2S	0.2694	0.2474	0.2510	0.2500	0.2492
	QML	0.2627	0.2442	0.2478	0.2518	0.2512
	FHLR	0.3716	0.3491	0.3523	0.2701	0.2508
Medium	SW	0.3329	0.3628	0.3556	0.2757	0.2803
	2S	0.2987	0.3013	0.2989	0.2488	0.2540
	QML	0.3070	0.3104	0.3089	0.2708	0.2793
	FHLR	0.3637	0.3628	0.3384	0.3404	0.2666
Large	SW	0.3066	0.2559	0.2711	0.2649	0.2441
	2S	0.2916	0.2778	0.2783	0.2803	0.2355
	QML	0.2942	0.2825	0.2837	0.2863	0.2577
	FHLR	0.3722	0.2881	0.3009	0.3003	0.2869

Table 3: RMSEs for the 3 data bases and the four factor extraction methods obtained by integrating sequentially the five factors in the leading equation (over the period Q1 2000 - Q4 2007).

4 Conclusions

From this application on the French GDP growth rate, we can conclude that complex dynamic models with strongly disaggregated data base do not necessarily lead to the best forecasting results. Indeed, the simple static Stock and Watson (2002a) approach with an aggregated data base of 20 series lead to comparable forecasting results when using a disaggregated data base of 140 series. Moreover, as a companion result, we empirically showed that the use of Bai and Ng (2002, 2007) tests would lead to unefficient forecasting results and that the inclusion of a higher number of factors improves the performances. Obviously, we do not claim that those results are general ones, but it would be interesting to continue this line of empirical research, with other data bases related to various countries, to check the robustness of our findings.

Further empirical research on this topic seems of great interest. For example, it would be interesting for practitioners to carry out a true-real time exercise taking the availability of data into account as well as vintage data. Other ways to forecast have been proposed in the literature on dynamic factor models, strongly associated with the factor extraction method (see Dias et al., 2008), for example by using the Kalman filter. It would be interesting to compare them with our global forecasting approach.

References

[1] Artis M., Banerjee A. and Marcellino M. (2005). Factor forecasts for the UK. Journal of Forecasting, 24, 135-172.

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

[3] Bai J. and Ng S. (2007). Determining the number of primitive shocks in factor models. Journal of Business and Economic Statistics, 25, 52-60.

[4] Banerjee A., Marcellino M. and Masten I. (2005). Leading indicators for the euro area inflation and GDP growth. Oxford Bulletin of Economics and Statistics, 67, 785-814.

[5] Barhoumi K., Benk S., Cristadoro R., Den Reijer A., Jakaitiene A., Jelonek P., Rünstler G., Ruth K. and Van Nieuwenhuyze C. (2008). Short-term forecasting of GDP using large data sets: A pseudo real-time forecast evaluation exercise. Journal of Forecasting, forthcoming.

[6] Boivin J. and Ng S. (2005). Understanding and comparing factor-based forecasts. International Journal of Central Banking, 1, 117-151.

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

[8] Breitung J. and Eickmeier S. (2006). Dynamic factor models. In H ubler O. and Frohn J. (eds.), *Modern Econometric Analysis*, Chap. 3, Springer.

[9] Breitung J. and Kretschmer U. (2005). Determining the number of dynamic factors in large macroeconomic panels. Mimeo, University of Bonn.

[10] Brisson M., Campbell B. and Galbraith J.W. (2003). Forecasting some low-predictability time series using diffusion indices. Journal of Forecasting, 22, 515-531.

[11] Camacho M. and Sancho I. (2003). Spanish diffusion indexes. Spanish Economic Review, 5, 173-203.

[12] Camba-Mendez G. and Kapetanios G. (2005). Forecasting euro area inflation using dynamic factor measures of underlying inflation. Journal of Forecasting, 25, 491-503.

[13] Cheung C. and Demers F. (2007). Evaluating forecasts from factor models for canadian GDP growth and core inflation. Working paper No 2007-8, Bank of Canada.

[14] Connor G. and Korajczyk R.A. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. Journal of Financial Economics, 15, 373-394.

[15] Connor G. and Korajczyk R.A. (1988). Risk and return in an equilibrium APT: Application of a new test methodology. Journal of Financial Economics, 21, 255-289.

[16] Connor G. and Korajczyk R.A. (1993). A test for the number of factors in an approximate factor model. Journal of Finance, 48, 1263-91.

[17] D'Agostino A. and Giannone D. (2007). Comparing alternative predictors based on large-panel factor models. CEPR Discussion Paper No 6564.

[18] Dempster A.P., Laird N. and Rubin D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B, 39, 1-38.

[19] Den Reijer A. (2005). Forecasting Dutch GDP using large scale factor models. DNB Working Paper No 28.

[20] Dias F, Pinheiro M. and Rua A. (2008). Forecasting using targeted diffusion indexes, Banco de Portugal, Working Paper No. 7/2008, May 2008.

[21] Diebold F.X. and Mariano R.S. (1995). Comparing predictive accuracy. Journal of Business and Economic Statistics, 13, 253-265.

[22] Doz C., Giannone D. and Reichlin L. (2006). A quasi maximum likelihood approach for large approximate dynamic factor models. CEPR Discussion Paper No. 5724.

[23] Doz C., Giannone D. and Reichlin L. (2007). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. CEPR Discussion Paper No 6043.

[24] Eickmeier S. and Ziegler C. (2008). How successful are dynamic factor models at forecasting output and inflation? A meta-analytic approach. Journal of Forecasting, 27, 3, 237-265.

[25] Engle R.F. and Watson M. (1981). A one-factor multivariate time series model of metropolitan wage rates. Journal of the American Statistical Association, 76, 774-781.

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

[27] Forni M., Hallin M., Lippi M. and Reichlin L. (2001). Coincident and leading indicators for the Euro area. Economic Journal, 111, 62-85.

[28] Forni M., Hallin M., Lippi M. and Reichlin L. (2003). Do financial variables help forecasting inflation and real activity in the euro area? Journal of Monetary Economics, 50, 1243-1255.

[29] Forni M., Hallin M., Lippi M. and Reichlin L. (2004). The generalized dynamic factor model: Consistency and rates. Journal of Econometrics, 119, 231-255.

[30] Forni M., Hallin M., Lippi M. and Reichlin L. (2005). The generalized dynamic factor model: One-sided estimation and forecasting. Journal of the American Statistical Association, 100, 830-840.

[31] Forni M. and Lippi M. (2001). The generalized dynamic factor model: Representation theory. Econometric Theory, 17, 1113–1141.

[32] Geweke J. (1977). The dynamic factor analysis of economic time series. In Aigner D.J. and Goldberger A.S. (eds.), *Latent Variables in Socio-Economic Models*, North-Holland, Amsterdam.

[33] Geweke J. and Singleton K.J. (1980). Maximum likelihood “confirmatory” factor analysis of economic time series. International Economic Review, 22, 37–54.

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

[35] Harvey D., Leybourne S. and Newbold P. (1997). Testing the equality of mean squared errors. International Journal of Forecasting, 13, 281-291.

[36] Jungbacker B. and Koopman S.J. (2008). Likelihood-based analysis for dynamic factor models. Working paper, VU University Amsterdam.

[37] Kapetanios G. (2004). A note on modelling core inflation for the UK using a new dynamic factor estimation method and a large disaggregated price index dataset. Economics Letters, 85, 63-69.

[38] Kapetanios G. and Marcellino M. (2004). A parametric estimation method for dynamic factor models of large dimensions. Queen Mary University of London Working Paper No 489.

[39] Marcellino M., Stock J. and Watson M. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. Journal of Econometrics, 135, 499-526.

[40] Quah D. and Sargent T.J. (1992). A dynamic index model for large cross-section. In Stock J. and Watson M. (eds.), *Business Cycle*, Univeristy of Chicago Press.

[41] Reichlin L. (2003). Factor models in large cross sections of time series. In Dewatripont M., Hansen L. and Turnovsky S. (eds.), *Advances in Econometrics, Theory and Applications, Eight World Congress of the Econometric Society*, Vol. III.

[42] Sargent T.J. and Sims C.A. (1977). Business cycle modeling without pretending to have too much a priori economic theory. In Sims C. et al. (eds.), *New Methods in Business Cycle Research*, Federal Reserve Bank ofMinneapolis, Minneapolis, MN.

[43] Schneider M. and Spitzer M. (2004). Forecasting Austrian GDP using the generalized dynamic factor model. OeNB Working Paper No 89.

[44] Schumacher C. (2007). Forecasting German GDP using alternative factor models based on large datasets. Journal of Forecasting, 26, 271-302.

[45] Stock J. and Watson M. (1989). New indexes of coincident and leading economic indicators. In Blanchard O.J. and Fischer S. (eds.), *NBER Macroeconomics Annual*, MIT Press.

[46] Stock J. and Watson M. (1999). Forecasting inflation. Journal of Monetary Economics, 44, 293-335.

[47] Stock J. and Watson M. (2002a). Macroeconomic forecasting using diffusion indexes. Journal of Business and Economic Statistics, 20, 147-162.

[48] Stock J. and Watson M. (2002b). Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association, 97, 1167-1179.

[49] Stock J. and Watson M. (2006). Forecasting with many predictors. In Elliott G., Granger C.W.J. and Timmermann A. (eds.), *Handbook of Economic Forecasting*, Vol. 1, North-Holland, Amsterdam.

[50] Watson M. 2003. Macroeconomic forecasting using many predictors. In Dewatripont M., Hansen L. and Turnovsky S. (eds), *Advances in Economics and Econometrics, Theory and Applications: Eighth World Congress of the Econometric Society*, Vol. III.

[51] White H. (1980). Maximum likelihood estimation of misspecified models. Econometrica, 50, 1-25.

Appendix A

In order to specify the number of factor Bai and Ng (2002, 2007) have suggested information criteria that can be used to estimate the number of factors consistently for static and dynamic model factors as n and T tend to infinity.

On the one hand, we determine the number of static factors r for SW by using the criterion IC_{p1} of Bai and Ng (2002) given by:

$$IC_{p1} = \ln\left(V(r,F)\right) + r.g(n,T), \qquad \text{with } g(n,T) = \left(\frac{n+T}{nT}\right)\ln\left(\frac{nT}{n+T}\right),$$

where $g(n,T)$ is a penalty function[9] and $V(r,F)$ measures the goodness-of-fit and is given by the following sum of squared residuals:

$$V(r,F) = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(X_t - \Lambda F_t\right)^2,$$

and depends on the estimates of the static factors and the number of factors. If the number of factors r increases, the variance of the factors also increases and the sum of squared residuals decreases. The estimated number of factors $\hat{r}$ is obtained from minimizing this information criterion, which reflects the trade-off between goodness-of-fit and overfitting.

On the other hand, the number of dynamic shocks q for dynamic principal component estimation of the factors and the state-space model is determined by information criterion proposed by Bai and Ng (2007). This criterion is obtained by taking the estimated static factors as given and then by estimating a VAR(p) model on these factors, where p is determined by the Bayesian information criterion (BIC). Then, they compute a spectral decomposition of the $(r \times r)$ residual covariance matrix $\widehat{\Gamma}_U$ and extract $\widehat{c}_j$ the j-th ordered eigenvalue, where $\widehat{c}_1 \succ \widehat{c}_2 \succeq ...\widehat{c}_r \succeq 0$. Then, for $k = 1,...,r-1$, they compute

$$\widehat{D}_k = \left(\frac{\widehat{c}_{k+1}}{\sum_{j=1}^{r}\widehat{c}_j}\right)^{1/2}$$

where each $\widehat{D}_k$ represents a measure of the marginal contribution of the respective eigenvalue, under the assumptions that $\widehat{\Gamma}_U = 0$ and that $c_k = 0$ for $k \succ q$. In practice, the set of admissible numbers of dynamic factors is chosen through the following boundary K=$\left\{k : \widehat{D}_k \prec m/\min\left[n^{\frac{2}{5}}, T^{\frac{2}{5}}\right]\right\}$. The number of dynamic factors is given by $\widehat{q} = \min\{k \in K\}$. In our application, we follow the Bai and Ng (2007) Monte Carlo results and we use $m = 1.0$.

[9] Bai and Ng (2002) proposed two others criteria, IC_{p2} and IC_{p3}, where the penalty function is defined as $g(n,T) = \left(\frac{n+T}{nT}\right)\ln C_{nT}^2$ and $(\ln C_{nT}^2/C_{nT}^2)$, respectively, with $C_{nT}^2 = \min\{n,T\}$.

Appendix B

1. Small data base

The small data base consists in 20 variables including:

A **Prices**: (1) Consumer price index (Insee); (2) Oil price Brent (Datastream).

B **Financial data**: (1) Rate of return on the long-term Government loans (monetary and financial statistics); (2) Treasury bonds with maturity of 13 weeks (monetary and financial statistics); (3) Reference rate of the regulated loans in housing (monetary and financial statistics); French stock index CAC40 (Datastream).

C **Soft data**: (1) Business sentiment indicator in industry (BdF); (2) Consumer sentiment indicator (Insee); (3) Services sentiment indicator (BdF); (4) Assessment of order-book levels (Eurostat); (5) Assessment of export order-book levels (Eurostat); (6) Production expectations for the months ahead (Eurostat); (7) Changes in retails sales (Insee).

D **Hard data**: (1) Household consumption in manufactured goods (Insee); (2) Industrial production index (Insee); (3) Exportations (Insee); (4) Importations (Insee); (5) Industrial car registrations (CCFA); (6) Declared housing starts (Ministry of Equipment).

2. Medium data base

For the medium data base, some soft data are disaggregated according to their various questions rather than using composite index. The disaggregated soft data are:

C1 **Business survey in industry**: (1) Order book by working week; (2) Total order book level; (3) Foreign order book level; (4) Change in total orders from previous month; (5) Change in delivery from previous month; (6) Change in foreign orders from previous month; (7) Change in production of finished goods from previous month; (8) Change in prices of finished goods from previous month; (9) Change in inventory of finished goods from previous month; (10) Change in staff levels from previous month; (11) Production forecast for the next month; (12) Inventory of finished goods forecast for the next month; (13) Inventory of commodities; (14) Inventory of finished goods; (15) Forecast staff level for the next month; (16) Capacity utilization rate.

C2 **Consumer confidence survey**: (1) Personal financial position past change; (2) Personal financial position outlook; (3) Living standards in France past change; (4) Living standards in France (outlook); (5) Timeliness of major purchases; (6) Personal financial position present level; (7) Future saving capacity; (8) Timeliness of saving; (9) Unemployment (outlook); (10) Prices (past change); (11) Prices (outlook).

C3 **Services activity survey**: (1) Changes in activity compared with the previous month; (2) Changes in prices compared with the previous month; (3) Changes in staff level compared with the previous month; (4) Cash flow situation; (5) Activity for the coming month; (6) Changes in price over the coming months; (7) Changes in staff level over the coming months.

3. Large data base

For large data base a sectorial disaggregation is applied for some data when possible.

- A11 **Consumer price index**. Each price data defined in A(1) is disaggregated as: (1) Agri-food; (2) Tobacco; (3) Manufactured goods; (4) Energy; (5) Services.
- C11 **Business survey in industry**. Each soft data defined in C1 is disaggregated as: (1) Intermediate goods; (2) Capital goods; (3) Automotive industry; (4) Consumer goods; (5) Agri-food industries.
- C71 **Changes in retails sales**. Each soft data defined in C(7) is disaggregated as: (1) New cars; (2) Old cars; (3) Textiles and clothing; (4) Furnitures; (5) Shoes; (6) Household electrical goods; (7) Electronics; (8) Hardware shops; (9) Watches and jewellers; (10) Agri-foods excluded meat; (11) Books and papers; (12) Meat.
- D11 **Household consumption**. Each hard data defined in D(1) is disaggregated as: (1) Cars; (2) Textile and leather; (3) Other manufactured goods; (4) Furnishing; (5) Household electrical; (6) Electronics.
- D12 **Industrial production index**. Each hard data defined in D(2) is disaggregated as: (1) Intermediate goods; (2) Capital goods; (3) Automotive industry; (4) Consumer goods; (5) Energy products.

Documents de Travail

229. Ph. Aghion, Ph. Askenazy, R. Bourlès, G. Cette and N. Dromel, “Education, Market Rigidities and Growth,” january 2009

230. G. Cette and M. de Jong, “The Rocky Ride of Break-even-inflation rates,” january 2009

231 E. Gautier and H. Le Bihan, “Time-varying *(S,s)* band models: empirical properties and interpretation,” January 2009

232. K. Barhoumi, O. Darné and L. Ferrara, “Are disaggregate data useful for factor analysis in forecasting French GDP ? ” February 2009

Pour accéder à la liste complète des Documents de Travail publiés par la Banque de France veuillez consulter le site : http://www.banque-france.fr/fr/publications/publications.htm

For a complete list of Working Papers published by the Banque de France, please visit the website: http://www.banque-france.fr/gb/publications/publications.htm

Pour tous commentaires ou demandes sur les Documents de Travail, contacter la bibliothèque de la Direction Générale des Études et des Relations Internationales à l'adresse suivante :

For any comment or enquiries on the Working Papers, contact the library of the Directorate General Economics and International Relations at the following address :

BANQUE DE FRANCE
49- 1404 Labolog
75049 Paris Cedex 01
tél : 0033 (0)1 42 92 49 55 ou 62 65 ou 48 90 ou 69 81
fax :0033 (0)1 42 92 62 92
email : thierry.demoulin@banque-france.fr
jeannine.agoutin@banque-france.fr
veronique.jan-antuoro@banque-france.fr
nathalie.bataille-salle@banque-france.fr