Forecasting French GDP with Dynamic Factor Models : a pseudo-real time experiment using Factor-augmented Error Correction Models

Stéphanie Combes, Catherine Doz

PARIS SCHOOL OF ECONOMICS
ECOLE D'ECONOMIE DE PARIS

WORKING PAPER N° 2018 – 28

Forecasting French GDP with Dynamic Factor Models : a pseudo-real time experiment using Factor-augmented Error Correction Models

Stéphanie Combes
Catherine Doz

JEL Codes: C22, E32, E37
Keywords :

PARIS-JOURDAN SCIENCES ECONOMIQUES
48, BD JOURDAN – E.N.S. – 75014 PARIS
TÉL. : 33(0) 1 80 52 16 00=
www.pse.ens.fr

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE – ECOLE DES HAUTES ETUDES EN SCIENCES SOCIALES
ÉCOLE DES PONTS PARISTECH – ECOLE NORMALE SUPÉRIEURE
INSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQUE – UNIVERSITE PARIS 1

Forecasting French GDP with Dynamic Factor Models : a pseudo-real time experiment using Factor-augmented Error Correction Models

Stéphanie COMBES [1] **and Catherine Doz** [2]

June 2018.

JEL Classification : C22, E32, E37

Abstract :

Dynamic Factor Models (DFMs) allow to take advantage of the information provided by a large dataset, which is summarized by a small set of unobservable latent variables, and they have proved to be very useful for short-term forecasting. Since most of their properties rely on the stationarity of the series, these models have been mainly used on data which have been differenciated to achieve stationarity. However estimation procedures for DFMs with I(1) common factors have been proposed by Bai (2004) and Bai and Ng(2004). Further, Banerjee and Marcellino (2008) and Banerjee, Marcellino and Masten (2014) have proposed to extend stationary Factor Augmented VAR models to the non-stationary case, and introduced Factor augmented Error Correction Models (FECM). We rely on this approach and conduct a pseudo-real time forecasting experiment, in which we compare short term forecasts of French GDP based on stationary and non-stationary DFMs. We mimic the timeliness of data, and use in the non-stationary framework the 2-step estimator proposed by Doz, Giannone and Reichlin(2011). In our study, forecasts based on stationary or non-stationary DFMs have a similar precision.

1. Institut National de la Statistique et des Etudes Economiques, France
2. Paris School of Economics and University Paris 1 - Panthéon Sorbonne

1 Introduction

Dynamic Factor Models (DFMs) have been extensively used since the 2000's and proved, in particular, to be very useful for short-term forecasting. These models allow to take advantage of the information provided by a large set of data, which is summarized by a small set of unobservable latent variables - the factors - taking into account the main comovements of the initial variables. The observable variables are then decomposed in two parts : one part (the common component) is a linear combination of the factors, and the other part (the idiosyncratic component) is specific to that variable. The factors can then be introduced in a forecast equation, which is called factor-augmented forecasting equation and which allows to use a large set of information when one wants to forecast a variable of interest for different forecast horizons. Moreover, as the estimation procedure can be adapted to handle missing values at the end of the sample due to publication lags, these models provide a flexible tool which can be used in order to get updated forecasts in real time, whenever new figures are released.

Since many of the DFMs' properties rely on the stationarity of the underlying series, these models have been mainly used in a stationary framework, meaning that the non-stationary series were transformed to achieve stationarity, and that the forecasted variable (for instance GDP) was taken in log-differences. However, the issue of estimating DFMs in a non-stationary framework has been addressed by several authors. Actually, the idea of representing a set of I(1) variables by a small set of I(1) factors is quite natural, since it is linked to the idea that I(1) variables may share common trends. Of course, this is just an intuitive presentation, since a factor model is characterized by the fact that the common factors are orthogonal to the idiosyncratic component (*i.e.* common factors are uncorrelated with idiosyncratic components at all leads and lags), whereas the classical common trend representation of a vector of I(1) series splits this vector into a random walk part (the common trends) and a stationary part which share the same innovation process. However, even though the common factor/idiosyncratic component decomposition does not coincide with the common trend/stationary component decomposition, it is clear that the idea of common trends is linked to the idea of common I(1) factors.

Dynamic factor models with I(1) common factors were first introduced by Bai (2004) and Bai and Ng (2004) in which common factors are I(1)and idiosyncratic components may be I(0) or I(1). In those papers, they propose to estimate the model using Principal Component Analysis (PCA) on the I(1) data, or on their first differences, and they provide consistency and asymptotic distribution results, as well as consistent criteria allowing to choose the appropriate number of factors. Building on this framework of I(1) DFMs, Banerjee and Marcellino (2009) and Banerjee, Marcellino and Masten (2014) have introduced Factor Augmented Error Correction models (FECM). These models are an extension to the non-stationary framework of Factor Augmented VAR (FAVAR) models, which have been first introduced by Bernanke, Boivin and Eliasz (2005) in the stationary framework. In FAVAR models, a few factors are added to a vector of interest variables, in order to add information to the VAR model without adding too many variables. These factors are usually supposed to represent the global state of the economy, and are obtained from a large set of macroeconomic variables. In the same way, Banerjee and Marcellino (2009) have proposed to add to a vector of I(1) variables of interest a small set of I(1) factors, associated to a large set of non stationary macroeconomic data, in order to add an informational content to the model. Cointegrating relations will generally exist between those factors and the variables of interest, so that the model can be written as an error correction model, which they call FECM. Bariggozzi, Lippi and Luciani (2017a, 2017b) provide a general framework for non stationary DFMs where the r factors are driven by $q < r$ dynamic shocks and can be cointegrated, and where the idiosyncratic components can be I(1) or I(0). In this case, the factors themselves follow a VECM whose parameters can be estimated by Johansen method, and they provide asymptotic results for the estimators.

In this paper, we run a pseudo-real time forecasting exercise based on stationary and non-stationary Dynamic Factor Models, and compare the associated results. This exercise is done for short forecast horizons (at most 7 months), in order to mimic forecasts which are made by the French Ministry of Finance. For the non stationary case, we rely on the FECM approach by Banerjee *et al.* (2009, 2014) and we also introduce factor-augmented forecast equations with an error-correction term. We compute pseudo-real time forecasts for French GDP for years 2000 to 2013 at a monthly frequency, which means that we mimic the conditions in which real forecasters were at each date, using (nearly) the same information as they had at that time[3]. In both the stationary and non-stationary cases, we estimate the factors using Doz, Giannone and Reichlin (2011) 2-step estimator, which allows to accommodate the jagged-edge data problem. We evaluate the quality of forecasts using RMSE of forecast errors. Our results don't show a real improvement of forecasts based on non-stationary DFMs over forecasts based on stationary DFMs for French GDP for the forecast horizons which we consider in this paper : the obtained RMSEs are similar in both cases. These results are comparable to Banerjee *et al.* (2014) results for Industrial Production at similar forecast horizons, however it should be noted that they obtain better results with FECM than with other methods, which is not our case. The fact that we don't obtain exactly the same kind of results may be linked to the fact that we are running a pseudo-real time experiment (meaning that we don't have a complete set of data at each date) or that we are forecasting a quarterly variable and not a monthly one, so that we have 3 times less observations when we estimate our forecasting equations.

The paper is organized as follows. In section 1, we recall the definitions and estimation methods for Dynamic Factor Models, both in the stationary and non-stationary frameworks. In section 2, we recall the techniques which are used to forecast an interest variable, using stationary DFM through a factor-augmented forecast equation, or using FECM when the framework is non-stationary. We also propose to use a factor-augmented forecast equation in the non stationary framework, in which an error correction term is introduced. In section 3, we describe our pseudo-real time experiment and give the main results. Section 4 concludes.

2 Dynamic factor models : stationary and non-stationary frameworks

2.1 Dynamic factor models in a stationary framework

Dynamic Factor Models have first been introduced in a stationary framework, and are mostly used in that framework. Even though these models are now well-known, we give below a very short reminder about their main features, in particular concerning the estimation methods, in order to stress the differences with the non-stationary framework in the next subsection. Detailed presentations of DFMs can be found in Bai and Ng (2008) and Stock and Watson (2010) and references therein.

In those models, each observation of a series x_i is split into two orthogonal components : one component (the common component) captures the bulk of cross-sectional comovements across series, and is driven by a few unobservable components - the common factors - and the other component (the idiosyncratic component) is uncorrelated with the factors at all leads and lags, and poorly cross-correlated with the other observed series. This decomposition is written as :

$$x_{it} = \lambda_i' F_t + e_{it}, \ \forall i = 1, \dots, n \tag{1}$$

3. In fact, it is not exactly the same information, because we don't take into account the fact that data are revised. In other words, we don't use "vintages" of data, but just take sub-samples of data which were available at the date where the forecast is computed.

where F_t is the vector of common factors and has size r with $r \ll n$, λ_i is the loadings vector, e_{it} is the idiosyncratic component, and $\text{Cov}(F_t, e_{is}) = 0 \ \forall (t,s)$ and $\forall i = 1, \ldots, n$.

Using matrix notations, the model is written as :

$$x_t = \Lambda F_t + e_t \tag{2}$$

where $x_t = (x_{1t}, \ldots, x_{nt})'$, $\Lambda = (\lambda_1, \ldots, \lambda_n)'$, $e_t = (e_{1t}, \ldots, e_{nt})'$.

Such a model relates x_t to the contemporaneous value of F_t only, but it is still a dynamic model, first because F_t is a dynamic process, and second because some components of F_t can be lagged values of other components. F_t and Λ are defined up to an invertible matrix, so that a normalization condition has to be imposed at the estimation stage.

The most commonly used estimation method is Principal Component Analysis (PCA), which has been in particular studied by Stock and Watson (2002), Bai (2003) or Bai and Ng (2002). For a given number k of factors, it amounts to solve the following problem

$$V(k) = \min_{\Lambda^k, F^k} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \lambda_i^{k\prime} F_t^k)^2 \tag{3}$$

subject to the normalization $F^{k\prime} F^k / T = I_k$ or to the normalization $\Lambda^{k\prime} \Lambda^k / N = I_k$. Under a set of standard assumptions, Stock and Watson (2002), or Bai (2003) have shown that this method provides consistent estimators of the loadings and of the factors when N and T grow to infinity. Without describing entirely this set of assumptions, let us just mention that, apart from the stationarity of the factors and idiosyncratic components, the main assumptions are, in both papers, that $\frac{\Lambda'\Lambda}{N} \longrightarrow \Sigma_\Lambda$ when $N \longrightarrow \infty$ (pervasiveness of the factors) whereas $\|Ve_t\|$ stays bounded when $N \longrightarrow \infty$. Further, Bai and Ng (2002) have proposed consistent information criteria, which allow to choose the number of factors.

Other methods can also be used to estimate the model for a given number of factors. In particular, Doz, Giannone and Reichlin (2011) have introduced a 2-step estimator which allows to consistently estimate the loadings and the factors, as well as the dynamics of the factors, when the factors vector follows a VAR model. This 2-step estimator has the advantage to be particularly well-suited for real-time estimation and forecasting. Indeed, it allows to cope very easily with the so-called "jagged-edge data problem", namely the fact that data are available with different timeliness so that, at a given date, some series are not available at the end of the sample (see Giannone, Reichlin and Small (2008) for an application of this method). The model can also be estimated by MLE (see Doz, Giannone, Reichlin(2012) and Bai and Li (2016)), but we won't discuss MLE here since, in the present paper, the model is estimated in the non-stationary case with the same 2-step estimator as in Doz *et al.* (2011).

The principle of this 2-step estimator can be described as follows (see Doz *et al.* (2011) for a complete presentation). The model is :

$$\begin{aligned} x_t &= \Lambda F_t + e_t \\ F_t &= \Phi_1 F_{t-1} + \cdots + \Phi_p F_{t-p} + \varepsilon_t \end{aligned} \tag{4}$$

and can easily be written in a state-space form.
It is supposed that $V\varepsilon_t = \Sigma$ and that $Ve_t = \Psi = \text{diag}(\psi_1, \ldots, \psi_n)$. This last assumption is generally not satisfied by the data (the idiosyncratic components may be weakly cross-correlated) however, provided that $\|Ve_t\|$ is bounded this 2-step method can be applied.

In the first step, a preliminary estimation $\hat{F}_t$ of the factors is obtained trough a principal component analysis of the balanced panel dataset (*i.e.* the largest sub-sample for which all series are available). The loadings and the Ψ matrix are then estimated through an OLS regression of the data on the estimated factors on the same sub-sample. A VAR(p) model is estimated with $\hat{F}_t$ playing the role of F_t. Doz *et al.* (2011) have shown that the obtained estimators $\hat{\Phi}_1, \ldots, \hat{\Phi}_p$ and $\hat{\Sigma}$ are consistent estimators of $\Phi_1, \ldots, \Phi_p$ and Σ.

In the second step, the Kalman filter and the Kalman smoother are used to re-estimate the factors, using the estimated values of the parameters which have been obtained in the first step, and using the complete set of available data. Indeed, as it is well known, the Kalman filter and smoother can easily deal with missing data. One way to do this is to set for instance :

$$Ve_{it} = \begin{cases} \psi_i \text{ if } x_{it} \text{ is available} \\ +\infty \text{ otherwise.} \end{cases} \tag{5}$$

Thus, the Kalman smoother provides a second estimation $\hat{F}_{t|T}$, based on the consistent estimators of the parameters which have been obtained in the first step, and on the information contained in the entire available dataset. Doz *et al.* (2011) show that $\hat{F}_{t|T}$ is a consistent approximation of F_t.

The advantages of this procedure are twofold : first, the factors can be estimated for the entire range of dates, even when there are missing data at the end of the sample, due to the jagged edge data problem ; second, the estimation of a VAR model for the factors can be subsequently used to forecast the factors and compute forecasts for other variables, based on the forecasted values of the factors.

2.2 Dynamic factor models in a non-stationary framework

As it has been mentioned before, DFMs are mostly used in a stationary framework and all the consistency properties have been first derived in this framework, under a standard set of assumptions. In particular, when the observed series are not stationary (generally $I(1)$), they are usually taken in differences (generally first differences), in order to get stationary series.

However, when series are $I(1)$ and cointegrated, it is well known that they share common trends (see e.g. Stock and Watson (1988)), so that it is natural to think of these common trends as sources of comovements. This intuition opens the way to the idea of integrated common factors, even though such a model does not rely on the same decomposition than the usual common trend/stationnary part decomposition as it has been put forward in the classical analysis of cointegrated series. Indeed, in the usual common trend/stationary part decomposition, the common trend and the stationary part are driven by the same innovation process, whereas in common factor models the common component and the idiosyncratic component are supposed to be orthogonal.

The extension of DFMs to the non-stationary framework has been proposed by Bai (2004) and Bai and Ng (2004). The most general case is given in Bai and Ng (2004). In this paper, the common factors and the idiosyncratic components can be stationary or non stationary. Each series is then decomposed as $x_{it} = \mu_i + \lambda_i' F_t + e_{it}$ (or $x_{it} = \mu_i + \beta_i t + \lambda_i' F_t + e_{it}$) where F_t is a $(r \times 1)$ vector which has r_1 non-stationary components and r_0 stationary components, and e_{it} is stationary or non-stationary. Thus, in this model, there are two possible sources of non-stationarity for a given series x_{it} : one source is the non-stationarity of some of the common

factors, and the other one is the possible non-stationarity of the idiosyncratic term. But the non-stationary common factors are the only source of non-stationarity which is common to a large number of series in the panel. As it will be mentioned below, the usual methods which are used to estimate factor models in a stationary framework cannot be used to estimate such a model. Indeed, in such a model, the loadings cannot be consistently estimated by OLS regression of the observed data on the estimated factors, since this may be tantamount to running a "spurious regression" (when e_{it} is non-stationary, x_{it} and F_t are not cointegrated).

In Bai (2004), on the contrary, idiosyncratic components are supposed to be stationary, and the non-stationary common factors are the only source for the non-stationarity of the series. In this case, under suitable assumptions, factor models' usual estimation methods can be applied.

2.3 Estimation of non-stationary dynamic factor models and determination of the number of factors

When common factors and idiosyncratic components may be I(1), Bai and Ng (2004) show that the model can be consistently estimated using the series in first differences. Indeed, if $x_{it} = \mu_i + \lambda_i' F_t + e_{it}$, then $\Delta x_{it} = \lambda_i' f_t + z_{it}$, where $f_t = \Delta F_t$ is stationary and has dimension r (with $r = r_0 + r_1$) and $z_{it} = \Delta e_{it}$ is stationary. Thus r can be consistently chosen according to Bai and Ng (2002) criteria, f_t can be estimated by PCA under standard assumptions, and the authors show that $\hat{F}_t = \sum_{s=2}^{t} \hat{f}_s$ consistently estimates F_t.

When idiosyncratic components are supposed to be I(0), Bai (2004) first considers the case where all the factors are I(1) (*i.e.* all the factors can be interpreted as common trends) and not cointegrated (otherwise there would be some I(0) factors which could be recovered through a rotation of the initial factors). The model he considers is then written as

$$x_{it} = \lambda_i' F_t + e_{it}, \text{ with } (1-L)F_t = u_t \tag{6}$$

where (u_t) is a stationary process, and each $(e_{it}), i = 1$ to n, is a stationary process. He extends to this framework the PCA estimation procedure used in the stationary factor models framework : for a given number k of factors, he estimates the model by solving

$$V(k) = \min_{\Lambda^k, F^k} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \lambda_i^{k\prime} F_t^k)^2$$

subject to the normalization $F^{k\prime} F^k / T^2 = I_k$ or to the normalization $\Lambda^{k\prime} \Lambda^k / N = I_k$. He shows that, under assumptions which generalize to the I(1) case the usual set of assumptions corresponding to the stationary case [4], the factors and the loadings are consistently estimated by this method.

He also introduces three information criteria, which extend to the non-stationary case the information criteria proposed by Bai and Ng (2002), and allow to consistently choose the appropriate number of factors (for a given maximum value $kmax$ of this number). These criteria can all be written as :

$$IPC_i(k) = V(k) + k\hat{\sigma}^2 \alpha_T g_i(N,T)$$

here $\alpha_T = T/(4\log(\log(T))$, $\hat{\sigma}^2 = V(kmax)$ and where $g_i(N,T), i = 1$ to 3 are suitable penalty functions.

4. In particular, the I(1) factors are supposed to satisfy the standard assumption : $\frac{1}{T^2} \sum_{t=1}^{T} F_t F_t' \to \int B_u B_u'$ when $T \to \infty$ with B_u a vector of Brownian motions.

Bai (2004) also considers the case where there are r_1 factors which are I(1) and r_0 factors which are I(0). In that case, the number of I(1) factors can be consistently estimated using the previous criteria, and the total number of factors can be consistently estimated using Bai and Ng (2002) criteria applied on the differenced data. Bai (2004) then proposes to estimate the I(1) factors from the r_1 first eigenvalues and eigenvectors of the matrix $\frac{1}{T^2}X'X$ and to estimate the I(0) factors from the $(r_1+1),\ldots,r_1+r_0$ largest eigenvalues and eigenvectors of $\frac{1}{T}X'X$.

Banerjee *et al.* (2014) mostly work under the same assumption as in Bai (2004), namely the assumption that the idiosyncratic component are stationary [5]. They use Bai(2004) criteria to fix the number of I(1) factors and use the same approach as Bai (2004) to estimate the factors. However, in the case where the model contains I(1) and I(0) factors, they also propose to estimate the stationary factors through a PCA of the residuals of the OLS regression of the I(1) series on the I(1) common factors (since the idiosyncratic components are supposed to be stationary, those residuals are stationary as well). At the forecasting stage, this case is associated to the model which they name as FECMc.

Two issues are worth noticing about the estimation of factor models with non-stationary data. First, even if Bai (2004) refers to Principal Component Analysis when he presents his estimation method, it should be noted that his method does not exactly resort to the usual meaning of PCA (where an eigen-decomposition of the covariance matrix or the correlation matrix of the data is done) but rather to a similar decomposition of the matrix of non central 2nd order moments. Indeed, the solution of the above optimization problem can be, for instance, obtained from the eigen-decomposition of the matrix $\frac{1}{T}X'X = \frac{1}{T}\sum_{t=1}^{T} x_t x_t'$, which is the matrix of non central 2nd order moments of the data. This is not innocuous since it means that the results depends on measurement units used for each series. Of course, this remark is also valid for PCA (the eigenvalues and eigenvectors of a covariance matrix are modified when measurement units of some series are changed) but, in the stationary framework, this problem can be easily overcome using centered and standardized data (in this case, PCA is run on the correlation matrix and not on the covariance matrix). In a non-stationary framework, considering centered and standardized data would not be interpretable, but the fact that the results depend on the measurement units is an issue which should probably be considered. In this paper, in order to address this issue we have decomposed the non central 2nd order moments matrix, but we have also considered the non central 2nd order moments matrix associated to $x_t - x_0$ instead of x_t in order to get rid of a part of this measurement unit problem (but there is still a scale effect left).

The second issue lies in the extraction of the factors. Bai (2004) proposes to extract the I(1) factors using the r_1 largest eigenvalues of $\frac{1}{T}X'X$ and the associated eigenvectors, and to extract the I(0) factors using the r_0 next eigenvalues and the associated eigenvectors. Indeed, the r_1 largest eigenvalues of $\frac{1}{T}X'X$, which are associated to I(1) factors, should theoretically diverge at rate NT, while the r_0 next eigenvalues, which are associated to the stationary factors, should diverge at rate N and the other eigenvalues should stay bounded : these properties theoretically allow to extract all factors simultaneously. However, in finite sample, it may happen that the eigenvalues don't display such a clear behaviour. For this reason, when we use PCA to estimate the factors, we have used the same approach as in Banerjee *et al.* (2014), namely we estimate the I(1) factors first, and we allow for the presence of I(0) factors underlying the residual of the OLS regression of the I(1) series on the I(1) common factors. However, as our database also contains I(0) series, we group those residuals with the I(0) series, and the stationary factors are extracted using this entire set of stationary series.

5. They also consider the case where the idiosyncratic component can be non-stationary, as in Bai-Ng (2004), but *ex post* checks lead them to consider that the stationarity of the idiosyncratic component is an assumption which is realized on their datasets.

2.4 Two-step estimator in a non-stationary framework

In this paper, we consider PCA as a first step for the estimation of the factors and we propose to estimate the stationary and the non-stationary factors using the 2-step estimator proposed by Doz *et al.* (2011). As we mentioned before, this estimator has indeed two advantages : first, all the available information can be used to estimate the factors, and the factors can be estimated for a larger range of dates than with PCA, since the method can be employed in case of jagged edge data ; second, this estimator can be used when the factors are supposed to follow a VAR model, and the estimated VAR model can be used when computing forecasts.

If, for instance, the factors are I(1) and not cointegrated [6], and if they are supposed to follow a VAR model, it is possible to cast the model into a state-space form, but the approach is slightly different from the approach which is used in the stationary case.

Indeed, if (F_t) is supposed to be a VAR(p) process, whose equation is : $\Phi(L)F_t = \mu + \varepsilon_t$, it is always possible to decompose $\Phi(L)$ as $\Phi(L) = \Phi(1) + (1-L)\Phi^*(L)$. If, further, (F_t) is supposed to be an I(1) non cointegrated process, then $\Phi(1)$ is necessarily equal to 0 (since its rank is equal to the cointegration rank). Then $\Phi(L)$ simplifies as :

$$\Phi(L) = (1-L)\Phi^*(L) = (1-L)(I - \Phi_1^* L - \dots \Phi_{p-1}^* L^{p-1}) \tag{7}$$

Thus, if G_t is defined by $G_t = (1-L)F_t$, the model can be written as :

$$\begin{aligned} x_t &= \Lambda F_t + e_t \\ F_t &= F_{t-1} + G_t \\ G_t &= \mu + \Phi_1^* G_{t-1} + \dots + \Phi_{p-1}^* G_{t-p+1} + \varepsilon_t. \end{aligned} \tag{8}$$

which can be easily cast in state-space form.

In particular, the state vector is then $\alpha_t = (F_t \; G_{t+1} \; \dots \; G_{t+1-p})'$, the transition equation can be written as :

$$\begin{pmatrix} F_t \\ G_{t+1} \\ G_t \\ \vdots \\ G_{t+3-p} \end{pmatrix} = \begin{pmatrix} 0 \\ \mu \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} I & I & 0 & \dots & 0 \\ 0 & \Phi_1^* & \Phi_2^* & \dots & \Phi_{p-1}^* \\ 0 & I & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & I & 0 \end{pmatrix} \begin{pmatrix} F_{t-1} \\ G_t \\ G_{t-1} \\ \vdots \\ G_{t+2-p} \end{pmatrix} + \begin{pmatrix} 0 \\ I \\ 0 \\ \vdots \\ 0 \end{pmatrix} \varepsilon_{t+1} \tag{9}$$

and the measurement equation can be written as :

$$x_t = \begin{pmatrix} \Lambda & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} F_t \\ G_{t+1} \\ \vdots \\ G_{t+1-p} \end{pmatrix} + e_t \tag{10}$$

with $Ve_t = \Psi = \mathrm{diag}(\psi_1, \dots, \psi_n)$ and $V\varepsilon_t = \Sigma$.

The 2-step estimator is computed as follows : in the first step, a preliminary estimation $\hat{F}_t$ of the factors is obtained trough a principal component analysis of the balanced panel dataset

6. The following representation can be extended straightforwardly to the case where there are I(1) factors and I(0) factors, and the initialization strategy which is presented below can be extended to this case as well.

(more precisely, as mentioned before, an eigen-decomposition of the uncentered second moments matrix). When the idiosyncratic components are supposed to be stationary, Bai (2004) shows that the estimated factors and loadings consistently estimate the true factors and loadings. A VAR(p) model is then estimated with $\hat{F}_t$ playing the role of F_t.

In the second step, the Kalman filter and the Kalman smoother are used to re-estimate the factors, using the estimated values of the parameters which have been obtained in the first step, and using the complete set of available data. Missing data can be treated using the identical trick as in the stationary case *i.e.* by defining :

$$Ve_{it} = \begin{cases} \psi_i \text{ if } x_{it} \text{ is available} \\ +\infty \text{ otherwise.} \end{cases}$$

As it is stated for instance in Harvey (1991) or Durbin and Koopman (2001), the Kalman filter and smoother can still be used in a non-stationary framework. The only change which is necessary, concerns the value which is chosen for the initial value of the covariance matrix P_1 of the state-vector at time $t = 1$. In a stationary framework, the unconditional covariance matrix of the state vector does not depend on time, and is a known function of the parameters, so that it can be taken as an initial value in the filter. In the non-stationary framework, the covariance matrix of the state vector goes to infinity with t. Consequently, a diffuse prior is generally adopted for the covariance matrix of the state vector at time $t = 1$. For instance, if the state vector is a random walk of size q, the covariance matrix is taken as $P_1 = \kappa I_q$ with $\kappa \to \infty$ (in practice κ is taken as an arbitrary large number). In our framework, the state vector cannot be taken as a random walk. However, the same kind of approach can be adopted, and it is possible to take a diffuse prior which is adapted to our VAR case (see appendix A1 for more details).

As in the stationary case, a new approximation $\hat{F}_{t|T}$ of F_t is obtained at the end of the second step : this estimation is based on the consistent estimations of the parameters which have been obtained in the first step, and uses all the information which is available in the dataset.

3 Forecasting with factors

3.1 Forecasting with a factor augmented equation in the stationary framework

In the stationary case, two approaches can be considered when a forecast is computed for a variable y_t using previously estimated factors (in order to have lighter notations we denote the estimated factors as F_t and not $\hat{F}_{t|T}$ in what follows). Both of them are based on a preliminary regression of the past values of y_t on the past values of the factors.

The first approach has been introduced by Stock and Watson (2002a, 2002b) and used e.g. by Boivin and Ng (2006) or Rünstler *et al.*(2009) among many others. In its basic form, when a forecast y_{T+h} has to be made, it relies on the OLS estimation of the following equation :

$$y_{t+h} = b + \beta' F_t + u_{t+h}, \ t = 1, \ldots, T-h \tag{11}$$

The forecast is then computed as $\hat{y}_{T+h|T} = \hat{b} + \hat{\beta}' F_T$.

In a more general form, lags of F_t and a vector W_t of explanatory variables (which can contain y_t and some of its lags) can be added, so that the estimated equation is :

$$y_{t+h} = b + \beta_0' F_t + \cdots + \beta_p' F_{t-p} + \gamma' W_t + u_{t+h}, \ t = p+1, \ldots, T-h \tag{12}$$

and the forecast is computed as $\hat{y}_{T+h|T} = \hat{b} + \hat{\beta}_0' F_T + \cdots + \hat{\beta}_p' F_{T-p} + \hat{\gamma}' W_T$.

The second approach is based on the dynamic behaviour of the factors, when the factors are supposed to follow a VAR model. It relies on a forecast of the factor itself, based on the estimated VAR model. This approach has been used in Giannone *et al.* (2008), and Banbura and Rünstler(2011) for instance. In the basic version, an OLS estimation of the contemporaneous link between the factors and y_t is first estimated :

$$y_t = c + \gamma' F_t + v_t. \tag{13}$$

Then a recursive forecast $F_{T+h|T}$ of the factors, based on the estimated VAR model, can be computed for any horizon h, and the forecast of y_{t+h} is finally computed as

$$\hat{y}_{T+h|T} = \hat{c} + \hat{\gamma}' F_{T+h|T} \tag{14}$$

In a more general version, lags of F_t and y_t can be added, so that the estimated equation is :

$$y_t = c + \sum_{k=0}^{p} \gamma_k' F_{t-k} + \sum_{k=1}^{p} \alpha_k y_{t-k} + v_t, \ t = p+1, \ldots, T-h \tag{15}$$

and forecasts are recursively computed as $\hat{y}_{T+h|T} = \hat{c} + \sum_{k=0}^{p} \hat{\gamma}_k' F_{T+h-k|T} + \sum_{k=1}^{p} \hat{\alpha}_k \hat{y}_{T+h-k|T}$.

3.2 Forecasting with FECM

As it has been said in the introduction, Banerjee *et al.* (2008,2014) have extended the FAVAR approach to the non-stationary framework and made a bridge between DFMs and cointegration analysis. Following the definition of FAVAR models in a stationary framework, they consider a VAR model which contains the I(1) interest variable y_t and the vector F_t of I(1) factors. When y_t and F_t are cointegrated, this VAR model will have an ECM representation, which they call Factor augmented Error Correction Model (FECM).

If β is a cointegrating matrix, such a model will have the following form

$$\begin{pmatrix} \Delta y_t \\ \Delta F_t \end{pmatrix} = \begin{pmatrix} \alpha_y \\ \alpha_F \end{pmatrix} \beta' \begin{pmatrix} y_{t-1} \\ F_{t-1} \end{pmatrix} + \Phi_1 \begin{pmatrix} \Delta y_{t-1} \\ \Delta F_{t-1} \end{pmatrix} + \cdots + \Phi_p \begin{pmatrix} \Delta y_{t-p} \\ \Delta F_{t-p} \end{pmatrix} + \begin{pmatrix} \varepsilon_{yt} \\ \varepsilon_{Ft} \end{pmatrix} \tag{16}$$

This kind of model can be defined when y_t is a one-dimensional interest variable, or when it is a vector (of small dimension) of interest variables, as it is the case in FAVAR models.

Such a model has two advantages : first, as in the FAVAR framework, it allows to take into account the information associated to a a large set of variables, which are summarized by F_t. Second, it allows to take into account the levels of the non-stationary variables, through the cointegration relations.

This model, as any VAR model, can be used for computing forecasts for the variable(s) of interest y_t. Banerjee *et al.* (2014) also introduce what they call FECMc, in which stationary factors can be added to the model, when they are relevant. They generally obtain good forecasting performances for both models.

3.3 Forecasting with a factor augmented equation including an error correction term

In this paper, we also explore an approach which mixes the two previous ones, and we compute the associated forecasts. Indeed, in the stationary framework, factor augmented forecast equations can be linked to the idea of FAVAR models. In the same way, one can associate to a FECM or a FECMc a factor augmented forecast equation using factors (and their lags) as explanatory variables, together with past values of the forecasted variable.

In the non-stationary framework, it is natural to include an error correction term inside such an equation. Indeed such an equation can be associated to an FECMc in the same way as a single error-correction equation can be associated to a VECM. In our framework, the interest variable y_t is unidimensional and as the I(1) factors are not cointegrated [7]. Thus, there is only one cointegration relation between the levels of the interest variable and of the I(1) factors, and this cointegrating relation corresponds to the error correction term. In the simplest case, when no lags are included, an analogous of equation (11) will thus have the following form :

$$\Delta y_{t+h} = c + \beta_0' F_{0t} + \beta_1' \Delta F_{1t} + \alpha \eta_{t-1} + u_{t+h},\ t = 1, \ldots, T-h \tag{17}$$

where F_{0t} is the vector of I(0) factors, F_{1t} is the vector of I(1) factors, $\eta_{t-1} = y_{t-1} - \delta' F_{1,t-1}$ is the error correction term at time $t-1$.

In the same way, an analogous of equation (13) will have the following form :

$$\Delta y_t = c + \gamma_0' F_{0t} + \gamma_1' \Delta F_{1t} + \alpha \eta_{t-1} + v_t,\ t = 1, \ldots, T. \tag{18}$$

Note that these two types of equations are natural extensions of equations (11) and (13) to the non-stationary framework and could be introduced without resorting to FECM or FECMc approach. Indeed if, as it can be expected, the variable of interest shares the common trends displayed by the other series, it must cointegrate with the I(1) factors which are proxies for the common trends. Equations (17) and (18) are thus ordinary error correction equations, dealing with stationary variables, and in which the levels of I(1) variables appear through a stationary cointegrating relation.

Under the assumption that there is a cointegrating relation between y_t and F_{1t} (which can be tested within the FECM or FECMc framework but can also be directly tested), Engle-Granger (1987) 2-step methodology can be applied : a super-consistent estimator of the cointegrating vector $(1\ -\delta)'$ can be obtained by OLS regression of y_t on F_{1t} and the coefficients of equations (17) and (18) can be estimated by OLS, using $\hat{\eta}_{t-1} = y_{t-1} - \hat{\delta}' F_{1,t-1}$ instead of η_{t-1}.

Then the forecasts associated to these equations are respectively computed as :

$$\begin{aligned} & \widehat{\Delta y}_{T+h|T} &=& \hat{c} + \hat{\beta}_0' F_{0T} + \hat{\beta}_1' \Delta F_{1T} + \hat{\alpha} \hat{\eta}_{T-1} \\ \text{and} \quad & \widehat{\Delta y}_{T+h|T} &=& \hat{c} + \hat{\gamma}_0' F_{0,T+h|T} + \hat{\gamma}_1' \Delta F_{1,T+h|T} + \hat{\alpha} \hat{\eta}_{T-1} \end{aligned}$$

where $F_{0,T+h|T}$ and $F_{1,T+h|T}$ can be obtained from the VAR model which has been separately estimated for the factors.
Of course, lags and other explanatory variables (including lags of the variable of interest) can be, as before, included in equations (17) and (18).

7. If the I(1) factors were cointegrated, then a rotation could be done and the initial I(1) factors could be replaced by a smaller number of I(1) factors plus a vector of I(0) factors obtained from the cointegrating relations. Here we suppose that a decomposition between I(1) and I(0) factors has already be done.

4 A Pseudo-real time forecasting experiment

4.1 Data, forecast horizons, and pseudo-real time design

In this paper, we use a database which is built on similar grounds as the Stock and Watson (2002) database, and which had been previously used in order to forecast French GDP with DFM, in a stationary framework (Bessec and Doz (2012)). The series are all monthly series and, in order to get a balanced dataset, the starting date has been fixed in January 1994.

This database is detailed in Appendix A2. It contains about a hundred variables and consists in survey data, real variables, nominal and financial variables, and international environment indicators like exchange rates, or some US and German economic indicators. Unit root tests have been run on all these series, and are mentioned in the table (in order to take the 2008 crisis into account, unit root tests have been run twice, using entire series as well as pre-crisis series).

We mimic forecasts which are made on a monthly basis for the next quarter (forecasting), the current quarter (nowcasting), and the previous quarter (backcasting). As a provisional figure for French GDP appears about 45 days after the end of the corresponding quarter, the backcasting exercise is done only once in each quarter, at the end of the first month of the quarter.

To be more precise, if we consider all forecasts made for quarter Q, and if we characterize the forecast horizon by the number of months between the date where the forecast is done, and the month when the provisional figure appears, forecasts horizons can be described as follows :

	Forecast	Forecast	Forecast	Nowcast	Nowcast	Nowcast	Backcast
Date of the forecast (Quarter & Month)	(Q-1) : M1	(Q-1) : M2	(Q-1) : M3	Q : M1	Q : M2	Q : M3	(Q+1) : M1
Horizon	7	6	5	4	3	2	1

The experiment is done mimicking real time conditions, meaning that, each time a forecast is computed, we only use the data which would have been available at that time in real life. For instance, if a series is published with a two months delay, this series is used with two missing values at the end of the sample. Such an exercise is however only a *pseudo* real time exercise since we don't use vintages of data. We use data which are available now, and we truncate them at the end of the sample, but we don't use data which were really available at the time when the forecast is computed. In other words, we don't take into account the fact that data are revised several times before final figures are published.

In our pseudo-real time experiment, forecasts are computed each month over the period 2001 to 2013, and they are based on data starting in January 1994 and ending in the month where the forecasts are computed. For each forecasting method under study, we compute the root mean square error associated to the whole set of forecasts.

Finally it is necessary to clarify how we implement the various models and forecasting methods which have been presented in the previous sections. Indeed, in our case, y_t is a quarterly series (here the quarterly GDP growth rate) and F_t is a monthly series. Two attitudes can then be adopted before estimating a forecast equation based on one of the above methods. One can interpolate y_t in order to get a monthly series, say y_t^m [8], estimate the forecast equation, using y_t^m and F_t as it has been estimated, in order to get monthly forecasts for y_t^m and finally

8. To avoid heavy notations, we use the same index t to describe the month t and the quarter to which this month belongs. This is of course a bit imprecise, but we think that the context should make things sufficiently clear for the reader.

compute a quarterly forecast for y_t based on those monthly forecasts of y_t^m. On the contrary, one can build a quarterly series F_t^q associated to the monthly series F_t and estimate the forecast equation using y_t and F_t^q. In this paper, we use this last approach, since it had been shown to give satisfying results in the stationary framework (see Bessec and Doz (2012)).

4.2 Different forecasting methods

4.2.1 Forecasting equations in the stationary framework

As we have seen in section 3, there are two types of forecasting equations which can be used when one wants to compute a forecast based on a factor-augmented regression equation (FAR). In our pseudo-real time exercise, with seven forecast horizons, for each month T where a forecast is made, if one uses the method which relies on the OLS estimation of (13), one only has to estimate this equation once, and to compute the 7 forecasts associated with the 7 horizons. On the other hand, if one uses equation (11), one has to run seven different estimations of (11), one for each horizon. We have used both approaches for our benchmark forecasting exercise, based on stationary data.

4.2.2 Forecasting with FECMc

In this paper, we have used the FECMc approach, rather than the FECM approach, since it was clear that stationary factors had to be taken into account in our data. We first estimate the I(1) factors using only the I(1) variables. Then, following Banerjee *et al.*(2014), we compute the residuals obtained in an OLS regression of the I(1) variables on the I(1) factors and we check that these residuals are stationary. We afterwards compute I(0) factors using both these residuals and the stationary variables of the initial dataset. Finally, all factors are estimated using the two-step approach which has been detailed before.

More precisely, if x_t is decomposed as $x_t = (X_{1t}\ X_{0t})'$, with X_{1t} the vector of non-stationary variables and X_{0t} the vector of stationary variables, and if we denote by F_{1t} the vector of non-stationary factors and F_{0t} the vector of stationary factors, the underlying factor model has the following form :

$$\begin{pmatrix} X_{1t} \\ X_{0t} \end{pmatrix} = \begin{pmatrix} \Lambda_1 & \Lambda_{10} \\ 0 & \Lambda_{00} \end{pmatrix} \begin{pmatrix} F_{1t} \\ F_{0t} \end{pmatrix} + \begin{pmatrix} E_{1t} \\ E_{0t} \end{pmatrix} \tag{19}$$

This factor model is estimated in 3 steps :

— in the first step $\hat{\Lambda}_1$ is obtained through a decomposition of the second moments matrix of X_{1t} on the balanced dataset

— in the second step $\hat{\Lambda}_0 = \begin{pmatrix} \hat{\Lambda}_{10} \\ \hat{\Lambda}_{00} \end{pmatrix}$ is obtained by PCA of the correlation matrix of $\begin{pmatrix} X_{1t} - \hat{\Lambda}_1 \hat{F}_{1t} \\ X_{0t} \end{pmatrix}$ on the balanced dataset [9]

— in the third step, Kalman filter and smoother are applied to get a new estimate of the factors based on the entire dataset.

As y_t is a quarterly variable, the FECMc corresponds to a VECM for the vector $(y_t\ F_{1t}^q\ F_{0t}^q)'$, where F_{it}^q is the quarterly value associated to F_{it} as in the stationary framework. As we mentioned before, since in our framework y_t is a univariate variable and the I(1) factors are not cointegrated, there is only one cointegration vector linking y_t to the I(1) factors. This cointegration vector, as well as the other parameters of the model, can be estimated using Johansen method and, for the estimated values of the parameters, recursive forecasts can be computed.

9. We have first check the stationarity of $X_{1t} - \hat{\Lambda}_1 \hat{F}_{1t}$.

Of course, as the factors are unobservable, the VECM is in fact estimated with $\hat{F}_{it|T}$ instead of F_{iT}, for $i = 0, 1$ and for the associated quarterly values.

4.2.3 Forecasting with a factor augmented equation including an error correction term

We have also computed forecasts based on the factor augmented equations including the error correction term which we have introduced in section 3. We have introduced lags of the factors and the interest variable in equations (17) and (18) which are thus respectively generalized as :

$$\Delta y_{t+h} = c + \sum_{k=0}^{p} \beta'_{0k} F^q_{0,t-k} + \sum_{k=0}^{p} \beta'_{1k} \Delta F^q_{1,t-k} + \gamma' W_t + \alpha \eta_{t-1} + u_{t+h},\ t = 1, \ldots, T-h$$

$$\text{and} \qquad \Delta y_t = c + \sum_{k=0}^{p} \gamma'_{0k} F^q_{0,t-k} + \sum_{k=0}^{p} \gamma'_{1k} \Delta F^q_{1,t-k} + \alpha \eta_{t-1} + \gamma' W_t + v_t,\ t = 1, \ldots, T.$$

where $\eta_{t-1} = y_{t-1} - \delta' F_{1,t-1}$ is the error correction term at time $t-1$ and W_t is a vector which can contain lags of y_t.

In practise, we have chosen $p = 1$, which is sufficient for our data (further lags are not significant). Let us also mention that, when one of these equations is used for backcasting (in practise, each first month of each quarter), the value which is forecasted is y_{T-1} and, in the equations, y_{T-1} is replaced by y_{T-2}. In particular, the error correction term is then η_{T-2} and not η_{T-1}.

4.3 The results

In our pseudo-real time forecasting exercise, we have compared the different forecasting approaches which have been previously mentioned : factor-augmented regression equation based on a stationary DFM, forecasts based on FECMc, and factor-augmented error-correction equation. Besides, for both types of factor-augmented forecasting equations, we have compared the results obtained with both kind of specifications : forecast based on the estimation of the simultaneous link between the factors and the forecasted variable (denoted by "method 1") v.s. forecasts based on different equations for the different horizons (denoted by "method 2"). Further, for the estimation of the non-stationary factors on the balanced dataset, we have considered the different ways of running PCA that we mentioned before (true PCA on centered and standardized data, true PCA on centered data, decomposition of the non centered 2nd moments matrix, decomposition of the 2nd moments matrix associated to $x_t - x_0$). Finally, we have compared the results obtained using the whole set of series, as well as different subsets of it.

For each method or dataset under study, the quality of the forecasts has been measured through the root mean square forecast error for each forecast horizon. Some general results can be stated. First, in the non-stationary case, the best results are generally obtained when the initialization is run using a decomposition of the non centered 2nd moments matrix, or of the 2nd moments matrix associated to $x_t - x_0$ rather than "true" PCA, as expected. Second, datasets which give the best forecasting results evolve a bit with the forecast horizon : roughly speaking, survey data are the most important ones for longer horizons, and the importance of real data increases when the forecast horizon decreases. Finally, for short horizons, the method which have been introduced (forecasting equation with an error correction term) seems to give more precise forecasts than the method based on a FECM, but both methods lead to a similar

quality of forecasts for longer horizons.

We propose here a selection of the results which we have obtained. In particular, for the non-stationary case, we present results associated to an initialization based on the decomposition of the 2nd moments matrix associated to $x_t - x_0$, since it overall give the best results.

Table : RMSE for different methods and horizons

	Stationary DFM		Non-stat DFM		
Horizon	FAR meth 1	FAR meth 2	FECM	FA−EC meth 1	FA−EC meth 2
-1	0.23	0.23	0.33	0.23	0.23
-2	0.25	0.24	0.32	0.28	0.28
-3	0.28	0.29	0.33	0.31	0.32
-4	0.30	0.30	0.35	0.32	0.34
-5	0.36	0.37	0.39	0.40	0.40
-6	0.42	0.40	0.37	0.40	0.38
-7	0.39	0.39	0.38	0.43	0.38

Notes :

- horizon is $-k$ when the forecast is computed with the information which is available k months before the first release of quarterly GDP
- FAR : forecast made with a Factor-Augmented regression (stationary framework)
- FA-EC : forecast made with a Factor-Augmented error correction equation (non-stationary framework)

It can be seen that, in our study, the best forecast quality is still obtained within the stationary framework. However a slight improvement is obtained with FECM for longer horizons, whereas univariate factor-augmented error correction equations don't seem to improve forecast quality.

Our results are not absolutely in line with to Banerjee *et al.* (2014) results for Industrial Production at similar forecast horizons, since they obtained better results with FECM than with other methods. Several facts may explain this difference. First, we are running a pseudo-real time experiment, so that we don't use a complete dataset at each date, which is not the case in Banerjee *et al.* (2014). Second, the fact that we forecast a quarterly series may deteriorate the quality of the forecast since, whichever method we use, we have three times less observations in the forecast equations. However, we have chosen to concentrate on GDP, which is of course a very important variable for forecasters. On the other hand, the fact that we obtain similar results for stationary and non-stationary frameworks show that the extension of Doz *et al.* (2011) 2-step method to a non-stationary framework is valid. This issue is addressed in another paper, which is currently in progress and will soon appear.

5 Conclusion

In this paper, we have used Banerjee *et al.* (2014) FECM framework in a pseudo-real time forecast experiment, where we compute monthly forecasts of French GDP for horizons running from one to seven months before its release. We have also proposed to use non-stationary dynamic factors inside univariate Error Correction forecasting equations. In both cases, the ragged-edge data problem has been addressed through an extension of Doz *et al.* (2011) 2-step method to a non-stationary framework. Our results don't show a real improvement when those forecasts are compared to forecasts computed with stationarized data and stationary factors. However, the precision of the forecasts obtained in both cases is similar, although a very slight improvement is observed for longer forecasts horizons with non-stationary factors.

References :

— Bai J. (2004) "Estimating cross-section common stochastic trends in nonstationary panel data", *Journal of Econometrics*,122, 137-183.

— Bai J. (2003) "Inferential Theory for Factor Models of Large Dimensions", *Econometrica*, 71(1), 135-171.

— Bai J. and K.Li (2016), "Maximum Likelihood Estimation and inference for Approximate Factor Models of High Dimension",*The Review of Economics and Statistics*, 2016, 98(2), 298-309.

— Bai J., Ng S. (2002), "Determining the Number of Factors in Approximate Factor Models", *Econometrica*, 70 (1), 191-221.

— Bai J., Ng S. (2004),"A PANIC Attack on Unit Roots and Cointegration", *Econometrica*, 72(4), 1127-1177.

— Bai J., and S. Ng (2007),"Determining the Number of Primitive Shocks in Factor Models", *Journal of Business and Economic Statistics* 25, 52-60.** supprimer ?**

— Bai J., and S. Ng (2008),"Large Dimensional Factor Analysis", *Foundations and Trends in Econometrics*, 3(2) : 89-163.

— Banerjee A., Marcellino M. (2009),"Factor-augmented Error Correction Models", *in* J.L. Castle and N.Shephard Eds *The methodology and practice of econometrics - a festschrift for David Hendry*, 227-254, Oxford University Press.

— Banerjee A., Marcellino M., Masten I. (2014),"Forecasting with Factor-augmented Error Correction Models",*International Jouranl of Forecasting*, 30, 589-612.

— Barigozzi, M., M. Lippi, and M. Luciani (2017a), "Dynamic factor models, cointegration, and error correction mechanisms", http ://arxiv.org/abs/1510.02399v3.

— Barigozzi, M., M. Lippi, and M. Luciani (2017b), "Non-Stationary Dynamic Factor Modelsfor Large Datasets", http ://arxiv.org/abs/1602.02398.

— Bernanke B.S., Boivin J., Eliasz P.(2005),"Measuring The Effects Of Monetary Policy : A Factor-Augmented Vector Autoregressive (FAVAR) Approach", *Quarterly Journal of Economics*, Vol. 120 (1,Feb), 387-422.

— Bessec, M. and Doz C. (2012), "Prévision à court terme de la croissance du PIB français à l'aide de modèles à facteurs dynamiques", *Economie et Prévision*, 199, 1-30.

— Doz C., Giannone D. and Reichlin L. (2011)"A two-step estimator for large approximate dynamic factor models based on Kalman filtering", *Journal of Econometrics* 164, 188-205.

— Durbin J. et S.J. Koopman (2001), *Time Series Analysis by State Space Methods*, Oxford Univ. Press.

— Engle R.F. et Granger C.W.J. (1987),"Cointegration and Error Correction Representation", Estimation and Testing, *Econometrica* 55, 251-276.

— Harvey A.C. (1991) *Forecasting, structural time series models and the Kalman filter*, Cambridge Univ. Press.

— Johansen S. (1995) *Likelihood-based inference in cointegrated vector-autoregressive models*, Oxford University Press.

— Onatski A. (2010),"Determining the number of factors from empirical distribution of eigenvalues", *The Review of Economics and Statistics* 92(4), 1004-1016.

— Stock J.H., M.W. Watson (1988),"Testing for Common Trends", *Journal of the American Statistical Association*, 83,no 404, 1097-1107.

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

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

— Stock J.H., M.W. Watson (2010),"Dynamic Factor Models",*in* Clements MP, Henry DF Oxford Handbook of Economic Forecasting, Oxford University Press.

Appendix

A1 : initialization of the Kalman filter in the non-stationary case

In what follows, we consider the same case as the one we use in the paper. We thus suppose that (F_t) is an I(1) non cointegrated process, and that (F_t) is a VAR(2) process, so that $\Phi(L)F_t = \mu + \varepsilon_t$, with $\Phi(L) = 1 - \Phi_1 L - \Phi_2 L^2$.
This case can be very easily extended to the general case where (F_t) is a VAR(p) process.

As $\Phi(L)$ can be decomposed as $\Phi(L) = \Phi(1) + (1-L)\Phi^*(L)$ and as $\Phi(1) = 0$ when (F_t) is not cointegrated, this can be written as

$$\Phi(L) = (1-L)(I - \Phi L)$$

since $\Phi^*(L)$ has degree 1 when $\Phi(L)$ has degree 2.

Thus, if G_t is defined by $G_t = (1-L)F_t$, the model can be written through the following state-space representation :

$$\begin{aligned} x_t &= \Lambda F_t + e_t \\ F_t &= F_{t-1} + G_t \\ G_t &= \mu + \Phi G_{t-1} + \varepsilon_t. \end{aligned}$$

If the state vector is $\alpha_t = (F_t \;\; G_{t+1})'$ the transition equation can be written as :

$$\begin{pmatrix} F_t \\ G_{t+1} \end{pmatrix} = \begin{pmatrix} 0 \\ \mu \end{pmatrix} + \begin{pmatrix} I & I \\ 0 & \Phi \end{pmatrix} \begin{pmatrix} F_{t-1} \\ G_t \end{pmatrix} + \begin{pmatrix} 0 \\ I \end{pmatrix} \varepsilon_{t+1}$$

and the measurement equation can be written as : $x_t = \begin{pmatrix} \Lambda & 0 \end{pmatrix} \begin{pmatrix} F_t \\ G_{t+1} \end{pmatrix} + e_t$.

When one wants to compute an initial value P_1 for the variance matrix of the state vector, one has to take into account the fact that VF_1 is not finite.

As $P_1 = \begin{pmatrix} VF_1 & Cov(F_1, G_2) \\ Cov(G_2, F_1) & VG_2 \end{pmatrix}$, we propose to choose the initial value using the following results :

- as it is usually done, VF_1 can be taken as $VF_1 = \kappa I$ with κ big enough ($\kappa \to \infty$)
- as (G_t) is a stationary VAR(1) process, $VG_t = \Gamma_G(0)$ can be computed as the solution of : $(I - \Phi \otimes \Phi) vec\Gamma_G(0) = vec\Sigma$, where $\Sigma = V\varepsilon_t$
- as, for any $H > 0$, $F_1 = F_{-H} + \sum_{s=-H+1}^{1} G_s$, $Cov(F_1, G_2)$ can be computed along the following way :

$$\begin{aligned} Cov(F_1, G_2) &= Cov(F_{-H}, G_2) + \sum_{s=-H+1}^{1} Cov(G_s, G_2) \\ &= Cov(F_{-H}, G_2) + \sum_{s=-H+1}^{1} \Gamma_G(s-2) \\ &= Cov(F_{-H}, G_2) + \sum_{s=-H+1}^{1} \Gamma'_G(2-s) \\ &= Cov(F_{-H}, G_2) + \sum_{h=1}^{H+1} \Gamma'_G(h) \end{aligned}$$

Since $G_t = \mu + \Phi G_{t-1} + \varepsilon_t$, we have $\Gamma_G(h) = \Phi^h \Gamma_G(0)$, and thus :

$$
\begin{aligned}
Cov(F_1, G_2) &= Cov(F_{-H}, G_2) + \sum_{h=1}^{H+1} \Gamma_G(0)\Phi'^h \\
&= Cov(F_{-H}, G_2) + \Gamma_G(0) \left(\sum_{h=1}^{H+1} \Phi^h \right)' \\
&= Cov(F_{-H}, G_2) + \Gamma_G(0) \left(\Phi \sum_{h=0}^{H} \Phi^h \right)'
\end{aligned}
$$

As (G_t) is stationary, the eigenvalues of Φ have modulus strictly smaller than 1, and $\sum_{h=0}^{H} \Phi^h \to (I - \Phi)^{-1}$ when $h \to \infty$.

Thus, if we suppose that $Cov(F_{-H}, G_2) \to 0$ when $H \to \infty$ we can take P_1 as :

$$
P_1 = \begin{pmatrix} \kappa I & \Gamma_G(0)(I - \Phi')^{-1}\Phi' \\ \Phi(I - \Phi)^{-1}\Gamma_G(0) & \Gamma_G(0) \end{pmatrix}
$$

Note that our approach is similar to the approach proposed by Durbin and Koopman(2001) but not identical, since we compute $Cov(F_1, G_2)$ which they take equal to 0. However, the two approaches can be shown to be asympotically equivalent.

A2 : The data.

We have used the following series :

Series	Freq.	Starting	Timeliness	Source
Survey Industry Past production	M	Apr. 76	M + 0	a
Industry Inventories	M	Apr. 76	M + 0	a
Industry Overall order books	M	Apr. 76	M + 0	a
Industry Export order books	M	Apr. 76	M + 0	a
Industry Personal production outlook	M	Apr. 76	M + 0	a
Industry General production outlook	M	Apr. 76	M + 0	a
Services Past activity	Q/M	1988	M + 0	a
Services Expected activity	Q/M	1988	M + 0	a
Services Expected demand	Q/M	1988	M + 0	a
Construction Past activity	Q/M	1975	M + 0	a
Construction Expected activity	Q/M	1975	M + 0	a
Construction Past employment	Q/M	1975	M + 0	a
Construction Order books	Q/M	1975	M + 0	a
Construction Productive capacity utilisation rate	Q/M	1975	M + 0	a
Retail trade General business outlook	M	1991	M + 0	a
Retail trade Past sales	M	1991	M + 0	a
Retail trade Orders	M	1991	M + 0	a
Retail trade Expected workforce	M	1991	M + 0	a
Consumers Personal− financial situation recent changes	M	1970	M + 0	a
Consumers Personal − financial situation expected change	M	1970	M + 0	a
Consumers Standard of living in France − recent changes	M	1970	M + 0	a
Consumers Standard of living in France − expected change	M	1970	M + 0	a
Consumers Right time to make major purchases	M	1970	M + 0	a
Manufacturing Demand − recent changes	Q	1976	Q+0	a
Manufacturing Demand − expected change	Q	1976Q2	Q+0	a
Services Past operating income	Q	1988	Q+0	a
Services Expected operating income	Q	1988	Q+0	a
Real Households New private vehicle registrations	M	1985	M + 1	a
Households Consumption of manufactured goods	M	1985	M + 1	a
Households Expenditure on durable goods	M	1985	M + 1	a
Households Expenditure on cars	M	1985	M + 1	a
Households Purchases of household durables	M	1985	M + 1	a
Households Consumption of textiles	M	1985	M + 1	a
Households Expenditure on other manufactured goods	M	1985	M + 1	a
Households Retail sales excluding cars	M	1995	M + 2	b
Households Unemployment rate	M	1983	M + 2	b
Households Unemployment rate for the under-25s	M	1983	M + 2	b
Households Job vacancies	M	1989	M + 2	c
Construction Total building starts	M	1994	M + 1	d
Construction Building starts (collective housing)	M	1994	M + 1	d
Construction Building starts (individual housing)	M	1994	M + 1	d
Construction Building starts (residential)	M	1994	M + 1	d
Construction Building permits (total)	M	1994	M + 1	d
Construction Building permits (collective housing)	M	1994	M + 1	d
Construction Building permits (individual housing)	M	1994	M + 1	d

Series	Freq.	Starting	Timeliness	Source
Industry IPI − total industry (BE)	M	1990	M + 2	a
Industry IPI − manufacturing (CZ)	M	1990	M + 2	a
Industry IPI − agricultural and food industries (C1)	M	1990	M + 2	a
Industry IPI − coking and refining (C2)	M	1990	M + 2	a
Industry IPI − electrical and electronic equipment (C3)	M	1990	M + 2	a
Industry IPI − cars (CL1)	M	1990	M + 2	a
Industry IPI − transport equipment excluding cars (CL2)	M	1990	M + 2	a
Industry IPI − other manufactured products (C5)	M	1990	M + 2	a
Industry IPI − mining and quarrying, energy and water supply (DE)	M	1990	M + 2	a
Industry IPI − construction (F)	M	1990	M + 2	a
Nominal Stock market CAC 40	M	1987	M + 0	e
Stock market SP 500	M	1980	M + 0	f
Stock market FTSE	M	1984	M + 0	g
Stock market DAX	M	1959	M + 0	h
Stock market Eurostoxx 50	M	1986	M + 0	i
Stock market Nikkei	M	1981	M + 0	j
Stock market Price Earnings Ratio (USA)	M	1954	M + 1	f
Stock market Volatility index (Vix)	M	1990	M + 0	k
Money M1	M	1980	M + 2	l
Money M2	M	1980	M + 2	l
Money M3	M	1980	M + 2	l
Money Loans (value)	M	1980	M + 2	l
Interest rate Mortgage interest rates	M	1978	M + 1	l
Interest rate 3−month interest rate	M	1970	M + 1	m
Interest rate 1−year interest rate	M	1960	M + 1	l
Interest rate 10−year interest rate	M	1989	M + 0	n
Interest rate Term structure (France)	M	1990	M + 0	o
Interest rate Term structure (USA)	M	1990	M + 0	o
Price Gold	M	1979	M + 0	p
Price Oil	M	1980	M + 1	m
Price Commodities	M	1980	M + 1	m
Price Consumer Price Index (CPI)	M	1990	M + 1	a
International Exchange rate euro/dollar	M	1978	M + 0	q
Exchange rate euro/sterling	M	1977	M + 0	q
Exchange rate euro/yen	M	1978	M + 0	q
Exchange rate euro/yuan	M	1978	M + 0	q
Exchange rate Nominal effective exchange rate for the euro	M	1979	M + 2	m
Exchange rate Real effective exchange rate for the euro	M	1979	M + 2	m
Germany Assessment of present business situation (IFO)	M	1991	M + 0	r
Germany Business expectations, next 6 months (IFO)	M	1991	M + 0	r
Germany Current economic situation (ZEW)	M	1992	M + 0	s
Germany Business expectations (ZEW)	M	1992	M + 0	s
Germany IPI − manufacturing	M	1991	M + 2	t
USA Retail sales (value)	M	1992	M + 1	u
USA IPI − manufacturing	M	1972	M + 1	v
USA Employment	M	1948	M + 1	w
USA Unemployment rate	M	1948	M +1	w
USA Manufacturing PMI	M	1948	M + 1	x

Notes :

— Q = quarterly ; M = monthly.

— IPI (C1) (or C2, C3, C5, in that order) represents the industrial production index in the following activities of the aggregated classification (NA) :
  C1 = Manufacture of food products and beverages,
  C2 = Manufacture of coke and refined petroleum products,
  C3 = Electrical and electronic equipment ; machine equipment,
  C5 = other manufacturing (excluding transport).
  CL1 = Manufacture of motor vehicles, trailers, and semi-trailers,
  CL2 = Manufacture of other transport equipment.

— Sources : a) INSEE ; b) Eurostat ; c) OECD ; d) French Ministry of Ecology ; e) Nyse Euronext Paris ; f) Standard & Poor's ; g) FTSE ; h) Frankfurt SE ; i) Financial Times ; j) OSE ; k) CBOE ; l) Bank of France ; m) IMF ; n) Daily press ; o) Global insight ; p) LBMA ; q) ECB ; r) IFO ; s) ZEW ; t) DESTATIS ; u) Census Bureau ; v) Federal Reserve Board ; w) BLS ; x) ISM.

— Timeliness : M+k for a series which is available k months after the end of month M.