CAHIERS

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

DOCUMENTS DE TRAVAIL DE LA DG TRÉSOR

Numéro 2014/03 – Décembre 2014

Prévisions de court terme du PIB : modèles à facteurs dynamiques et non stationnarité

Stéphanie COMBES
Catherine DOZ

TRÉSOR
DIRECTION GÉNÉRALE

PRÉVISION DE COURT TERME DU PIB : MODÈLES À FACTEURS DYNAMIQUES ET NON STATIONNARITÉ

Stéphanie COMBES
Catherine DOZ

* **Stéphanie COMBES** était en poste à la Direction Générale du Trésor du Ministère des Finances et des comptes Publics et du Ministère de l'Économie, de l'Industrie et du Numérique (France)
stephanie.combes@insee.fr

* **Catherine DOZ** est en poste à l'Université Paris 1 Panthéon-Sorbonne et Paris School of Economics
catherine.doz@univ-paris1.fr

TRÉSOR
DIRECTION GÉNÉRALE

Table des matières

Résumé

Les modèles à facteurs sont de plus en plus utilisés pour la prévision de court terme du PIB par les banques centrales et les grands organismes internationaux. Ces modèles permettent en effet de résumer l'information apportée par un grand nombre de variables en un petit nombre de variables latentes appelées facteurs. Sous leur forme dynamique, ces modèles permettent de rendre compte des évolutions conjointes des indicateurs qui les constituent. De plus, en recourant à des techniques d'estimation appropriées, il est possible de résoudre les problèmes posés par les différences de délais de publication des variables utilisées. De cette façon, il n'est pas nécessaire de développer différents modèles en fonction de la date de prévision et des variables disponibles à cette date-là.

Dans une précédente étude, avaient été examinées les performances en prévision de ces modèles pour la prévision du taux de croissance du PIB français sur des horizons courts, en utilisant une base constituée d'une centaine de variables, pour la plupart stationnaires ou rendues stationnaires par une transformation. Or il est bien connu que la plupart des séries macroéconomiques doivent être considérées comme non stationnaires. Depuis la fin des années 1980 et l'introduction de la notion de cointégration, on sait que, lorsqu'on travaille avec des variables non stationnaires, il peut être utile de prendre en compte l'information sur les relations d'équilibre de long terme entre ces variables. Classiquement, la prise en compte de ces relations de cointégration se fait dans le cadre d'un modèle de petite taille, appelé modèle à correction d'erreur. Ce type de modèle ne permet d'utiliser qu'un petit nombre de variables pour calculer cette prévision. Il peut donc sembler utile de combiner les avantages des MFD (la prise en compte d'un grand nombre de variables) et des modèles à correction d'erreur (la prise en compte des relations de long terme ou de cointégration) pour la prévision de court terme. La mise en œuvre d'un tel modèle nécessite un certain nombre d'aménagements qui en accroissent la complexité. Ceci pourrait expliquer que les performances obtenues ne soient pas meilleures.

Abstract

Factor models have received increasing interest from central banks and international organizations due to their ability to forecast short-term activity. In these models, a large set of variables is summarized by a small set of unobservable variables, referred to as factors. This allows to take advantage of the information provided by the large set of initial variables. In their dynamic form, these models can also take into account the comovements of the underlying variables. Moreover, the estimation procedure can be adapted to handle the missing values at the end of the sample due to publication lags. Thus, it is no longer necessary to develop separate models for different forecasting dates, even if information sets differ from one date to another.

In a recent paper, the accuracy of these models in forecasting French GDP growth rate over short horizons was investigated, using a large data set of a hundred of variables. A special attention was paid to consider only stationary variables in the set of data. However, it is well known that most macroeconomic variables are non-stationary. Since the end of the eighties and the introduction of cointegration concept, we know that it can be useful to take into account the long term equilibrium relationships between these variables, that are referred to cointegration relationships in the modelisation. Error correction models have been developed in this purpose. These models only allow a small amount of variables, so it seemed relevant for short-term forecasts to try to combine advantages from factor models (their ability to resume a lot of information) and error correction models that integrate cointegration relationships. Implementing such a model requires a lot of modifications. As a consequence, performance may be penalized by the complexity.

Introduction

La prévision de croissance du PIB à court terme (à l'horizon du trimestre passé avant sa publication, du trimestre en cours ou du trimestre suivant) fournit les conditions initiales à la prévision de plus long terme, et constitue donc le point de départ de la construction des budgets. Par ailleurs, le bureau de l'analyse conjoncturelle à la DG Trésor communique régulièrement ses estimations de la croissance à court-terme au ministre, ce qui constitue un élément d'information primordial pour l'appréciation en temps réel de la situation économique du pays.

Parmi les outils de prévision du bureau, sont utilisées des approches dites « directes », qui donnent, de façon agrégée, une première appréciation du dynamisme de l'économie. Elles servent de complément précieux lors de la réalisation des prévisions macro-sectorielles, car elles fournissent un repère qui permet d'ajuster les postes pour lesquels peu de données sont disponibles ; d'autre part, elles permettent d'atteindre des horizons de prévision plus éloignés.

Entre autres approches directes, les *modèles à facteurs dynamiques (MFD)*, qui permettent de résumer l'information apportée par un grand nombre de variables en un petit nombre de variables latentes, appelées facteurs, se sont révélés très performants sur la période récente. Ils présentent l'avantage sur les autres modèles (étalonnages) d'intégrer toute l'information disponible au fil de l'eau, ce qui en fait des outils très opérationnels et fiables.

Cependant, la littérature sur les modèles à facteurs concerne principalement la modélisation de variables stationnaires, et c'est donc dans ce cadre stationnaire qu'ont été développés les modèles actuellement utilisés. Or il est bien connu que la plupart des séries macroéconomiques doivent être considérées comme non stationnaires et qu'elles ne sont stationnaires qu'après différenciation : une grande partie des variables utilisées pour construire ces modèles ont donc été prises en différences premières (lorsque les variables sont en logarithme, cela signifie que l'on utilise le taux de croissance de ces variables).

Mais depuis la fin des années 1980 et l'introduction de la notion de cointégration, on sait que lorsqu'on travaille avec des variables non stationnaires, il peut être utile de prendre en compte l'information sur les relations d'équilibre de long terme entre ces variables, qui se traduisent par l'existence de relations de cointégration entre elles. Classiquement, la prise en compte de ces relations de cointégration se fait dans le cadre d'un modèle de petite taille, appelé modèle à correction d'erreur. Ce type de modèle se révèle utile pour la prévision de court terme, même s'il ne permet d'utiliser qu'un petit nombre de variables pour calculer cette prévision : les travaux de F. Charpin (2011) montrent par exemple qu'un modèle simple faisant intervenir un petit nombre de variables et une relation de cointégration entre le PIB et les variables consommation et exportations, donne des performances similaires à celles d'un modèle à facteur.

Il peut donc sembler utile de combiner les avantages des MFD (la prise en compte d'un grand nombre de variables) et des modèles à correction d'erreur (la prise en compte des relations de long terme ou de cointégration). De fait, les travaux réalisés sur données américaines par Banerjee, Marcellino et Marsten (2010), semblent montrer que la prise en compte de ces relations d'équilibre au sein de modèles à facteurs dynamiques est susceptible d'améliorer la précision de la prévision de court terme.

Conserver l'avantage du modèle à facteurs dynamique, tout en intégrant la possibilité d'avoir des relations de cointégration entre variables en niveau, est donc une voie naturelle à explorer pour améliorer les outils de prévision de court terme.

1. Les modèles à facteurs dynamiques : cadre stationnaire et non stationnaire

L'utilisation des modèles à facteurs dynamiques a pris un essor considérable depuis le début des années 2000, notamment pour l'utilisation de ces modèles en prévision. Ces modèles permettent en effet de prendre en compte une grande quantité d'information et d'utiliser les données en temps réel, au fur et à mesure de leur publication[1]

Plus précisément, on suppose dans ces modèles que les variables observées peuvent être décrites en fonction d'un petit nombre de variables latentes inobservables, appelées facteurs, dont la dynamique rend compte de l'essentiel de la dynamique commune aux variables. Une fois ces facteurs estimés, il est possible d'écrire une équation de prévision de la variable d'intérêt en la régressant sur ces facteurs.

Ces modèles ont été principalement développés et utilisés dans un cadre stationnaire : les variables utilisées pour extraire les facteurs, ainsi que les facteurs eux-mêmes, sont alors supposés stationnaires. Ceci impose de différencier les variables étudiées lorsqu'elles sont non stationnaires (ce qui est le cas de la plupart des séries macroéconomiques). Cependant certains auteurs ont récemment développé et utilisé ces modèles dans un cadre non stationnaire, c'est-à-dire dans un cadre où les variables étudiées et les facteurs sous-jacents peuvent être non stationnaires.

Après avoir rappelé rapidement les propriétés principales des modèles à facteurs dynamiques dans le cadre stationnaire, nous présentons ci-dessous les travaux principaux qui ont été consacrés à l'étude de ces modèles dans le cadre non stationnaire, ainsi que le cadre retenu pour l'utilisation de ces modèles pour la prévision.

1.1 Le modèle à facteur dynamique : cadre stationnaire

1.1.1 Le modèle

Le principe des modèles à facteurs dynamiques consiste à supposer qu'un vecteur $x_t = (x_{1t}, \ldots, x_{nt})'$ de variables observées peut être correctement représenté sous la forme :

$$x_t = \Lambda_0 f_t + \cdots + \Lambda_s f_{t-s} + e_t$$

avec $f_t = (f_{1t}, \ldots, f_{qt})'$ un vecteur de taille q (q<<n) de variables inobservables appelées facteurs (ou facteurs communs) et e_t un vecteur de variables inobservables, non corrélées aux facteurs, et appelées composantes idiosyncratiques.

On suppose souvent que la dynamique de f_t est régie par un modèle VAR :

$$f_t = \sum_{i=1}^{p} A_i^0 f_{t-i} + \varepsilon_t \qquad (1)$$

avec (ε_t) un bruit blanc.

On montre facilement que ce type de modèle peut être aussi écrit sous la forme suivante :

$$x_t = \Lambda F_t + e_t \qquad (2)$$

$$F_t = A\, F_{t-1} + \underbrace{B\varepsilon_t}_{\zeta_t} \qquad (3)$$

[1] Une présentation détaillée est fournie par Bessec et Doz (2011), *qui a été repris dans Bessec et Doz (2012)* ».

Dans cette écriture, le vecteur F_t contient f_t et ses *J* premiers retards (avec $J = max(s, p-1)$) : $F_t = (f_t', \ldots, f_{t-J}')'$, Λ et *A* sont des matrices qui s'expriment de façon simple en fonction des matrices $\Lambda_0, \cdots, \Lambda_s, A_1, \ldots A_p$ et $B = (I_q, 0 \ldots 0)'$. F_t est ici un vecteur de taille *r,* avec $r = (J+1)q$: on dit que l'on a un modèle dans lequel le nombre de facteurs statiques est *r* (les facteurs statiques sont les composantes de F_t), et le nombre de facteurs dynamiques est *q* (les facteurs dynamiques sont les composantes de f_t).

Le cadre le plus fréquemment étudié est celui des MFD approchés, dans lequel les composantes du vecteur e_t peuvent être corrélées entre elles, de façon contemporaine ou au cours du temps, mais où la part de la dynamique des variables observables qui est liée aux composantes idiosyncratiques est négligeable devant la part liée aux facteurs, lorsque le nombre de variables étudiées *n* tend vers l'infini. Ceci se traduit par un certain nombre d'hypothèses que nous ne détaillerons pas ici, mais il importe de préciser que l'on suppose en particulier toujours que les valeurs propres de la matrice $\Lambda'\Lambda$ tendent vers l'infini avec *n* (le plus souvent on suppose qu'elles croissent à la même vitesse que *n*).

1.1.2 Estimation des facteurs et détermination de leur nombre

Diverses méthodes d'estimation de ces modèles ont été proposées dans la littérature (voir Bai et Ng (2008b), Stock et Watson (2010), ou Barhoumi, Darné et Ferrara (2012) pour un survol complet de ces méthodes). La méthode la plus couramment employée consiste à estimer le modèle (2) par Analyse en Composantes Principales (ACP) : la matrice Λ et le facteur statique F_t sont obtenus à partir des *r* vecteurs propres associés aux *r* plus grandes valeurs propres de la matrice de variance-covariance des données observées. Sous les hypothèses usuellement retenues dans la spécification du modèle à facteurs approchés, on montre que l'ACP permet d'obtenir des estimateurs convergents des paramètres du modèle et une approximation des facteurs qui converge vers leur vraie valeur lorsque le nombre *n* des séries étudiées et le nombre *T* des observations tendent vers l'infini.

D'autres méthodes d'estimation ont été proposées pour permettre la prise en compte de la dynamique des facteurs. En particulier Doz, *et al* (2011) ont proposé une méthode d'estimation en deux étapes basée sur le filtre de Kalman, qui a été utilisée dans les travaux de Bessec et Doz (2012) et qui est depuis employée par le bureau de l'analyse conjoncturelle de la Direction générale du trésor. Elle présente l'avantage de s'adapter facilement au cas de valeurs manquantes qui est un des problèmes importants auquel le conjoncturiste doit faire face.

L'approche la plus couramment retenue pour déterminer le nombre *r* de facteurs statiques à retenir dans le modèle est celle qui a été proposée par Bai et Ng (2002). Elle consiste à choisir le nombre *r* qui minimise un critère d'information, dont la forme générale est :

$$PC(k) = V(k, F_k) + k\, g(n,T) \text{ ou } IC(k) = \ln\left(V\left(k, \widehat{F_k}\right)\right) + kg(n,T),$$

où *g* est une fonction de pénalité (pour laquelle les auteurs proposent 3 spécifications possibles) et où $V(k, \hat{F}_k) = \frac{1}{nT}\sum_{t=1}^{T}\sum_{t=1}^{T}\hat{e}_{it}^2$ est la variance empirique moyenne des composantes idiosyncratiques obtenue lorsque le modèle est estimé avec *k* facteurs. Ces auteurs ont aussi proposé (*cf.* Bai et Ng (2007)) des critères permettant de sélectionner le nombre *q* de facteurs dynamiques à retenir.

1.1.3 Utilisation en prévision

Lorsque, comme c'est le cas ici, la variable à prévoir est une variable trimestrielle (notée y_t), alors que les données utilisées, et donc aussi les facteurs, sont à valeurs mensuelles, la prévision requiert généralement une trimestrialisation des facteurs. La prévision repose ensuite sur une régression de y_t (ici le taux de croissance trimestriel du PIB) sur les valeurs contemporaines et/ou passées des facteurs trimestrialisés. Deux approches sont alors généralement utilisées : si l'on note $f_{i,t}^{Q}$ la valeur trimestrialisée estimée du *i*ème facteur à la date *t,* la première approche consiste à estimer par MCO le modèle :

$$y_{t+h} = \sum_{i=1}^{r} \delta_i f_{i,t}^{Q} + \varepsilon_{t+h} \qquad t = 1,\ldots,T-h \tag{4a}$$

et à calculer ensuite la prévision de y_{T+h} à la date *T* en utilisant la formule suivante :

$$\hat{y}_{T+h|T} = \sum_{i=1}^{r} \hat{\delta}_i f_{i,T}^{Q} \tag{4b}$$

La deuxième approche est spécifiquement liée au cadre dynamique et utilise l'estimation de la dynamique des facteurs qui est obtenue lorsqu'on estime le modèle à facteurs. En effet, si les facteurs vérifient un modèle de la forme $f_t = \sum_{i=1}^{p} A_i^0 f_{t-i} + \varepsilon_t$, il est possible d'obtenir de façon récursive une prévision $f_{T+h|T}$ de f_{T+h} à la date *T* en utilisant les valeurs estimées des matrices A_i^0 et des facteurs. On peut ensuite trimestrialiser les prévisions obtenues et déterminer une prévision $f_{T+h|T}^{Q}$ du vecteur des facteurs trimestrialisés. On estime alors par les moindres carrés ordinaires l'équation reliant le taux de croissance du PIB aux facteurs trimestrialisés qui lui sont contemporains $f_{i,t}^{Q}$:

$$y_t = \sum_{i=1}^{r} \delta_i f_{i,t}^{Q} + \varepsilon_t \qquad t = 1,\ldots,T \tag{5a}$$

et la prévision de y_{T+h} à la date *T* est obtenue en utilisant la formule suivante :

$$\hat{y}_{T+h|T} = \sum_{i=1}^{r} \hat{\delta}_i f_{i,T+h|T}^{Q} \tag{5b}$$

1.2 Le modèle à facteurs dans un cadre non stationnaire

1.2.1 Introduction : notion de tendance commune

Les notions de tendances communes à un ensemble de séries non stationnaires, et de cycles communs à un ensemble de séries, sont apparues dans la littérature à la fin des années 1980 et au début des années 1990, parallèlement au développement de la notion de cointégration entre séries non stationnaires. Deux articles ont été à l'initiative de ces notions : l'article de Stock et Watson (1988) qui définit les tendances communes à partir d'une généralisation à un vecteur de séries de la décomposition que Beveridge et Nelson (1981) avaient introduite pour les séries univariées, et celui de Engle et Kozicki (1993) qui introduit la notion de cycle commun.

Dans la décomposition de Stock et Watson (SW), comme dans celle de Beveridge et Nelson (BN), la notion de tendance commune doit être cependant entendue dans un sens particulier puisqu'il s'agit en fait d'une marche aléatoire vectorielle commune aux séries étudiées. Plus

précisément, si $(1-\mathrm{L})\mathrm{x_t} = \mu + C(L)\epsilon_t$ est la décomposition de Wold du processus vectoriel $(\mathrm{x_t})$ (dont les composantes sont les séries étudiées), la décomposition de BN-SW est la suivante :

$$\mathrm{x_t} = T_t + C_t \text{ avec } (1-\mathrm{L})\mathrm{T_t} = \mu + C(1)\epsilon_t \text{ et } C_t = C^*(L)\epsilon_t = \frac{C(L)-C(1)}{1-L}\epsilon_t$$

Lorsque $(\mathrm{x_t})$ est cointégré avec un rang de cointégration égal à *r* et lorsque β est une matrice de cointégration[2], on montre (cf Stock et Watson (1988) ou Johansen (1995)) que C(1) est une matrice de rang $n-r$ et que la marche aléatoire T_t peut s'écrire comme une fonction linéaire d'une marche aléatoire W_t de dimension $n-r$. Les composantes de cette marche aléatoire constituent ce que l'on appelle les tendances communes au sens de BN-SW. L'intérêt de cette décomposition réside dans le fait de décomposer la série initiale $(\mathrm{x_t})$ en la somme de $n-r$ marches aléatoires, sur lesquelles l'impact d'un choc ϵ_t est permanent et d'une composante stationnaire, sur laquelle l'impact d'un choc est transitoire. Sa spécificité tient cependant d'une part au fait que les tendances communes sont spécifiées comme des marches aléatoires, ce qui est restrictif, et d'autre part au fait que les innovations des deux composantes sont identiques, à une matrice près, puisque l'innovation de la composante stationnaire est ϵ_t et celle de la marche aléatoire commune est $C(1)\epsilon_t$.

Cette décomposition sert souvent de référence aux auteurs qui se sont intéressés aux modèles à facteurs dans un cadre non stationnaire (voir par exemple Escribano et Pena (1994), Banerjee et Marcellino (2008), ou Banerjee, Marcellino et Masten (2010)). Néanmoins, elle ne peut en réalité pas être directement utilisée pour écrire un modèle à facteurs, du fait de la corrélation entre la composante non stationnaire et la composante stationnaire. En revanche, l'idée sous-jacente à cette décomposition selon laquelle, lorsqu'un processus vectoriel non stationnaire est cointégré, il existe des tendances non stationnaires qui sont communes à l'ensemble des séries, peut être très naturellement reprise dans le cadre des modèles à facteurs. En effet, dans le cadre de ces modèles, comme on l'a vu précédemment dans le cadre stationnaire, on suppose qu'une grande partie des comouvements des séries peut être prise en compte par la dynamique d'un nombre restreint de facteurs communs : il est alors très naturel de supposer que certains de ces facteurs communs sont non stationnaires et prennent en compte les chocs qui ont un impact à long terme sur l'économie, tandis que d'autres sont stationnaires et prennent en compte les chocs qui n'ont qu'un impact transitoire sur l'économie.

1.2.2 Le modèle à facteurs dans le cadre non stationnaire : généralités

L'extension des modèles à facteurs au cadre non stationnaire a été proposée par Bai et Ng (2004) et par Bai (2004).

L'article de Bai et Ng (2004) fournit le cadre le plus général : dans cet article, les facteurs communs peuvent être stationnaires ou non stationnaires et les composantes idiosyncratiques peuvent aussi être stationnaires ou non stationnaires. Chaque série se décompose sous la forme $\mathrm{x_{it}} = \mu_i + \lambda_i' F_t + e_{it}$ (ou $\mathrm{x_{it}} = \mu_i + \beta_i t + \lambda_i' F_t + e_{it}$ éventuellement) et le vecteur F_t se décompose en r_0 composantes stationnaires et r_1 composantes non stationnaires. La non stationnarité de chaque série $\mathrm{x_{it}}$ peut donc provenir à la fois de la non stationnarité de certains facteurs communs et de la non-stationnarité de la composante idiosyncratique e_{it}, mais les facteurs communs non stationnaires rendent compte des sources de non-stationnarité communes à l'ensemble des séries non stationnaires, alors que les composantes idiosyncratiques ne peuvent être communes qu'à un nombre fini de séries. Comme on le verra ci-dessous la présence éventuelle de composantes idiosyncratiques non stationnaires a un impact très important sur la méthode d'estimation des facteurs, et sur la détermination de leur nombre. En effet, les méthodes proposées dans le cas stationnaire ne peuvent pas s'étendre

[2] Rappelons qu'un processus vectoriel $(\mathrm{x_t})$ est cointégré lorsqu'il existe une combinaison linéaire stationnaire $\mathrm{b'x_t}$. On dit alors que b est un vecteur de cointégration. L'ensemble des vecteurs de cointégration forme un sous espace de dimension *r*. On dit que β est une matrice de cointégration lorsque ses vecteurs colonnes forment une base du sous-espace de cointégration.

à ce cadre, en particulier parce qu'une régression des variables initiales sur les facteurs estimés serait ce qu'on appelle une régression fallacieuse (*spurious regression*) : lorsque e_{it} est non stationnaire, x_{it} et les composantes non stationnaires de F_t ne sont pas cointégrées, la régression de x_{it} sur F_t ne peut pas être valide.

Dans l'article de Bai (2004) au contraire, les composantes idiosyncratiques sont supposées stationnaires, et les facteurs non stationnaires sont la seule source de non-stationnarité des séries étudiées. Dans ce cas, sous des hypothèses qui généralisent celles qui sont faites dans le cadre stationnaire, on peut mettre en œuvre des méthodes d'estimation des facteurs et de détermination de leur nombre qui sont similaires à celles qui sont employées dans le cas stationnaire.

1.2.3 Estimation des facteurs et détermination de leur nombre dans le cas où les composantes idiosyncratiques peuvent être I(1)

Dans l'article de Bai et Ng (2004), l'idée principale est que les séries différenciées (qui sont stationnaires) vérifient elles aussi un modèle à facteurs, dans lequel les facteurs et les composantes idiosyncratiques sont stationnaires et vérifient les hypothèses usuelles. En effet, on peut écrire[3] :

$$\Delta x_{it} = \lambda_i' f_t + z_{it} \text{ avec } f_t = \Delta F_t \text{ et } z_{it} = \Delta e_{it}$$

Ce modèle peut être estimé de façon convergente par ACP, et les critères de détermination du nombre total de facteurs $r = r_0 + r_1$ définis par Bai et Ng (2002) peuvent être employés.

On peut ensuite définir $\hat{F}_t = \sum_{s=2}^{t} \hat{f}_s$ et $\hat{e}_{it} = \sum_{s=2}^{t} \hat{z}_{is}$

Les auteurs montrent alors les résultats suivants :

- l'hypothèse de non stationnarité de e_{it} peut être testée au moyen d'un test ADF usuel sur $\hat{e}_{it}$;
- dans le cas où il n'y a qu'un seul facteur commun, l'hypothèse de non stationnarité de ce facteur peut être testée au moyen d'un test ADF usuel sur $\widehat{F_t}$;
- dans le cas où le vecteur F_t est de taille $r > 1$, le nombre de facteurs non stationnaires peut être déterminé de façon séquentielle en utilisant un test fondé sur les valeurs propres de la matrice d'autocorrélation d'ordre 1 de $\widehat{F_t}$;
- $\widehat{F_t}$ est un estimateur convergent de F_t

1.2.4 Estimation des facteurs et détermination de leur nombre dans le cas où les composantes idiosyncratiques sont I(0)

Lorsque les composantes idiosyncratiques sont stationnaires, l'article de Bai (2004) propose de travailler directement avec les séries en niveau, non différenciées, et non stationnaires. Nous nous concentrons ici sur les sections de l'article dans lesquelles on suppose que les facteurs sont intégrés d'ordre 1 et non cointégrés.

On suppose donc que : $x_{it} = \lambda_i' F_t + e_{it}$ et $(1 - L)F_t = u_t$, avec (u_t) et (e_{it}) des processus stationnaires.

Le terme F_t représente bien les tendances communes aux x_{it} mais, contrairement à ce qui est fait dans la décomposition de BN-SW, on n'impose pas ici à ces tendances communes d'être des marches aléatoires et e_{it} n'est pas corrélé à F_t.

Sous des hypothèses qui généralisent au cadre non stationnaire les hypothèses usuellement faites sur les modèles à facteurs (*cf.* par exemple Bai et Ng (2002) pour une description de ces hypothèses usuelles), l'auteur montre que, lorsqu'il y a *r* tendances communes, les facteurs communs et les *loadings* λ_i peuvent être estimés de façon convergente par les *r* premières composantes principales obtenues en faisant une ACP sur les séries en niveau.

[3] On ne présente ici que le cas sans trend temporel.

Par ailleurs, l'auteur propose 3 nouveaux critères de détermination du nombre de facteurs (ici, comme on l'a dit, on suppose que tous les facteurs sont non stationnaires). Ces critères ont la même forme que ceux qui ont été proposés par Bai et Ng (2002), mais ils reposent sur un terme de pénalité différent, qui tient compte du caractère non stationnaire des facteurs. La forme générale de ces critères est donc la suivante : $IPC(k) = V\left(k, \widehat{F_k}\right) + k\alpha_T \, g(n,T)$ où les trois spécifications proposées pour la fonction de pénalité *g* sont les mêmes que dans le cas stationnaire.

Si on note k le nombre de facteurs, il s'agit donc de minimiser l'un des trois critères suivants :

$$IPC_1(k) = V\left(k, \widehat{F_k}\right) + k\,\widehat{\sigma_2}\,\alpha_T \left(\frac{N+T}{NT}\right)\log\left(\frac{NT}{N+T}\right)$$

$$IPC_2(k) = V\left(k, \widehat{F_k}\right) + k\,\widehat{\sigma_2}\,\alpha_T \left(\frac{N+T}{NT}\right)\log(\min(N,T))$$

$$IPC_3(k) = V\left(k, \widehat{F_k}\right) + k\,\widehat{\sigma_2}\,\alpha_T \left(\frac{N+T-k}{NT}\right)\log(NT)$$

avec $\alpha_T = T/[4\log\log(T)]$, $\hat{\sigma}_2 = V(kmax)$ en pratique.

1.2.5 Autres approches

D'autres auteurs se sont intéressés à la présence éventuelle de facteurs non stationnaires. Nous mentionnons ici les deux articles qui nous semblent les plus intéressants. Concernant la détermination du nombre de facteurs, il faut en effet citer l'article important d'Onatski (2010). Si l'on compare cet article avec le reste de la littérature sur les modèles à facteurs, il faut d'abord préciser que cet auteur travaille sous un ensemble d'hypothèses différent de celui qui est habituellement utilisé. Pour résumer, on peut dire qu'il fait des hypothèses plus générales sur la matrice des *loadings* (les valeurs propres de la matrice $\Lambda'\Lambda$ sont supposées tendre vers l'infini avec *n*, mais pas nécessairement toutes à la même vitesse), des hypothèses plus générales sur les facteurs (qui peuvent être stationnaires ou non stationnaires) et des hypothèses spécifiques sur la forme de la composante idiosyncratique, qui sont plus restrictives que les hypothèses usuelles. Sous cet ensemble d'hypothèses, cet auteur propose une procédure de test permettant de déterminer le nombre de facteurs, basée sur l'écart entre les valeurs propres consécutives de la matrice de variance-covariance des données. Les simulations présentées par l'auteur semblent montrer que cette procédure donne de bons résultats, mais on peut regretter que ces simulations ne concernent que le cas où les facteurs sont stationnaires[4].

Un autre article qui nous semble devoir être mentionné est celui de Barigozzi et *al.* (2013). Ces auteurs considèrent eux aussi le cas de modèles à facteurs dans lesquels certains facteurs sont stationnaires et d'autres sont non stationnaires. Ils explorent en particulier la distinction entre modèles à facteurs statiques et dynamiques en étendant à ce cadre ce qui avait été introduit dans le cadre stationnaire (rappelons qu'un modèle à *q* facteurs dynamiques admet aussi une représentation statique dans laquelle interviennent $r > q$ facteurs). Considérant que l'hypothèse de stationnarité de la composante idiosyncratique est une hypothèse trop forte, ces auteurs appliquent la procédure de Bai et Ng (2004) pour déterminer le nombre de facteurs statiques et estimer ces facteurs. Ils proposent ensuite une procédure fondée sur l'analyse de la matrice de densité spectrale pour étudier les facteurs dynamiques, mais sans avoir étudié cette procédure de façon approfondie à ce stade.

[4] Nous avons implémenté le test pour pouvoir l'utiliser dans notre étude (*cf.* annexe 3).

1.3 Modèles FECM et application à notre étude

1.3.1 Présentation générale

Dans le cadre stationnaire, Bernanke, Boivin et Eliasz (2005) avaient introduit les modèles FAVAR (Factor Augmented VAR models) dans lesquels on estime un modèle VAR sur un vecteur regroupant d'une part les séries d'intérêt, et d'autre part un facteur commun (ou un petit nombre de facteurs communs) estimé(s) à partir d'un grand nombre d'autres séries, et susceptible(s) de rendre compte du reste de l'économie. Comme les modèles VAR ne peuvent être estimés que sur un petit nombre de variables, la présence du (ou des) facteur(s) commun(s) a l'avantage d'élargir l'ensemble d'information pris en compte dans la modélisation VAR.

Cependant, lorsque l'on modélise sous forme VAR un vecteur de séries non stationnaires, et lorsque ces séries sont cointégrées, on peut montrer (voir par exemple Johansen (1995)) que la spécification correcte du modèle est une représentation appelée "vectorielle à correction d'erreur" (VECM) qui décrit les liens entre les valeurs contemporaines des séries différenciées (interprétables comme des taux de croissance lorsque les séries étudiées sont en logarithme), leurs valeurs retardées, et les écarts aux relations de long terme (les relations de cointégration) mesurés à la période précédente. Ainsi, lorsque le vecteur y_t est non stationnaire mais cointégré, avec une matrice de cointégration β, on montre que, si y_t admet une représentation VAR d'ordre p de la forme $y_t = \mu + \sum_{k=1}^{p} A_k y_{t-k} + \epsilon_t$, alors il admet aussi une représentation VECM, équivalente à la représentation VAR initiale, et qui a la forme suivante :

$$\Delta y_t = \mu - \alpha\beta' y_{t-1} + \sum_{k=1}^{p-1} \Phi_k \Delta y_{t-k} + \epsilon_t$$

Cette représentation a l'intérêt d'être pertinente sur le plan économétrique, puisqu'elle ne fait intervenir que des variables stationnaires (les séries différenciées et les relations de cointégration) tout en conservant, *via* les relations de cointégration, l'information sur les niveaux des variables. À l'inverse, un modèle VAR qui ne tiendrait compte que des relations dynamiques entre les variables différenciées, serait mal spécifié et ne tiendrait pas compte de toute l'information disponible.

Se fondant sur cette idée d'une perte d'information lorsqu'on travaille exclusivement sur des variables stationnarisées par différenciation, Banerjee et Marcellino (2008), puis Banerjee, Marcellino et Masten (2010) ont repris la même idée que Bernanke, Boivin et Eliasz (2005) mais en se plaçant dans le contexte des modèles VAR non stationnaires. Ils ont alors introduit les modèles FECM (Factor augmented Error correction Models) qui permettent, comme dans les modèles FAVAR, de tenir compte d'un ensemble élargi d'information *via* des facteurs extraits d'un ensemble large de données, et qui permettent aussi, comme dans les modèles VECM, de prendre en compte l'information sur les niveaux des variables, *via* les relations de cointégration. Si l'on dénote par y_t le vecteur (de petite taille) des variables d'intérêt, et par f_t le vecteur (de petite taille aussi) des facteurs extrait d'un large ensemble de données, un tel modèle aura donc la forme suivante :

$$\begin{pmatrix} \Delta y_t \\ \Delta f_t \end{pmatrix} = \begin{pmatrix} \alpha_y \\ \alpha_f \end{pmatrix} \beta' \begin{pmatrix} y_{t-1} \\ f_{t-1} \end{pmatrix} + \Phi_1 \begin{pmatrix} \Delta y_{t-1} \\ \Delta f_{t-1} \end{pmatrix} + \cdots + \Phi_p \begin{pmatrix} \Delta y_{t-p} \\ \Delta f_{t-p} \end{pmatrix} + \begin{pmatrix} \epsilon_{y_t} \\ \epsilon_{f_t} \end{pmatrix}$$

1.3.2 Aspects méthodologiques des travaux de Banerjee et *al.*

Banerjee et *al.* (2008, 2010) proposent deux démarches pour l'extraction des facteurs non stationnaires. Dans la première, les facteurs I(1) sont extraits en utilisant seulement les séries I(1) de la base alors que, dans la deuxième, les auteurs utilisent aussi les valeurs cumulées des séries stationnaires[5]. Ils indiquent cependant que les résultats finaux sont qualitativement très semblables, quelle que soit la démarche retenue.

[5] Les valeurs cumulées d'une série x_{it} sont définies par $X_{it} = \sum_{s=1}^{t} x_{is}$

Le nombre de facteurs I(1) est déterminé en utilisant les critères proposés par Bai (2004), et le nombre de facteurs I(0) est déterminé en utilisant les critères de Bai et Ng (2002). Les facteurs I(1) sont calculés en utilisant la méthode préconisée par Bai (2004), c'est-à-dire en utilisant l'ACP sur les séries en niveau. Les auteurs ne précisent cependant pas s'ils ont contrôlé la stationnarité des composantes idiosyncratiques, qui est pourtant une condition *sine qua non* de validité des résultats de Bai (2004). Les facteurs stationnaires sont estimés par composantes principales sur la base complète, dans laquelle les séries non stationnaires sont remplacées par leurs différences premières.

Cependant, le nombre de facteurs I(1) et le nombre de facteurs I(0) obtenus n'étant en pratique pas compatibles dans l'exemple étudié, les auteurs suggèrent que les variables I(1) en niveau peuvent avoir à la fois des tendances communes et des cycles communs : on serait donc en présence d'un cas où les variables I(1) auraient à la fois des facteurs communs I(1) et des facteurs communs I(0), ce qui est tout à fait plausible. Pour traiter ce cas, les auteurs estiment d'abord les facteurs I(1), puis calculent les résidus de la régression des variables I(1) sur les facteurs I(1), et mènent ensuite une ACP sur ces résidus, supposés stationnaires, pour en extraire un ou des nouveau(x) facteur(s) commun(s) stationnaire(s) qui peuvent éventuellement être ajoutés aux facteurs stationnaires précédents.

Les facteurs stationnaires et non stationnaires sont ensuite inclus dans le modèle FECM, dont le nombre de retards est déterminé par application du critère de Hannan Quin. Ce modèle est estimé comme un modèle VECM usuel et, sous les hypothèses de Bai (2004), le fait que les facteurs inobservables soient remplacés par les facteurs estimés ne change pas les propriétés des estimateurs obtenus. Le modèle est ensuite utilisé pour effectuer des prévisions des variables d'intérêt y_t.

1.4 Les problèmes posés par les travaux sur les modèles à facteurs non stationnaires

Dans les articles qui traitent de la présence de facteurs non stationnaires que nous venons de citer, il nous semble qu'il existe un certain nombre de problèmes qui ne sont pas ou peu abordés. Ces problèmes et l'existence de plusieurs façons de les traiter, nous amèneront, dans la partie appliquée de notre travail, à faire un certain nombre d'essais dont nous comparerons les résultats afin d'établir quelle est la méthode qui nous semble la plus pertinente à employer.

La façon dont il faut traiter les séries non stationnaires lorsqu'on les analyse par ACP constitue le premier de ces problèmes. En effet, il est bien connu que les résultats de l'ACP ne sont pas invariants par changement d'échelle (c'est-à-dire par changement des unités de mesure de chacune des séries qui constituent le vecteur de séries analysé). Ce problème est passé sous silence dans l'ensemble de la littérature sur les modèles à facteurs (stationnaires ou non stationnaires) alors qu'il nous semble tout de même mériter qu'on s'y attarde. Dans le cadre stationnaire, certains auteurs font le choix de centrer et réduire les variables étudiées, pour les ramener toutes à une même échelle (c'est ce qui a été fait par exemple dans Doz et *al.* (2011,2012) ou dans Bessec et Doz (2012)). Dans ce cas, c'est en fait la matrice de corrélation empirique entre les variables qui est analysée et non pas leur matrice de variance-covariance.

Cependant, s'il est facile de donner une interprétation au fait de centrer et réduire une variable x_t dans un cadre stationnaire, puisque la moyenne empirique $\bar{x}$ est un estimateur de l'espérance commune des x_t, et la variance empirique un estimateur de la variance commune des x_t, une opération analogue sur une variable non-stationnaire semble beaucoup plus difficilement justifiable (rappelons que si le processus (x_t) est non stationnaire, l'espérance de x_t n'est pas nécessairement constante, et la variance de x_t augmente avec t). De fait, dans leurs articles, Bai (2004) et Bai et Ng (2004) utilisent la matrice des moments non centrés d'ordre deux des séries prises en logarithme[6], et non l'analyse par composantes principales sur la matrice de corrélation. Cependant, l'impact du choix des unités de mesure n'est pas du

[6] Ceci n'est pas précisé dans les articles mais a été confirmé par les auteurs.

tout étudié par ces auteurs, et cela nous semble poser problème lorsque les séries étudiées sont de nature très différente et qu'elles sont mesurées dans des unités très différentes.

Un deuxième problème réside dans le fait que la littérature existante ne permet pas de définir clairement quelle est la méthode d'estimation à retenir lorsqu'on est en présence simultanément de facteurs intégrés d'ordre 1 et de facteurs stationnaires. Cette question est abordée dans Bai (2004) mais elle est traitée d'une façon qui, comme nous l'avons indiqué précédemment, ne nous semble pas précise. Mentionnons cependant l'idée générale de la procédure d'estimation par ACP qu'il propose : si l'on a r_1 facteurs I(1) et r_0 facteurs I(0), alors les r_1 plus grandes valeurs propres de la matrice de variance-covariance de x_t tendront vers l'infini en nT voire $n\frac{T}{\log\log T}$ si suffisamment de conditions sont remplies d'après Bai (2004), les r_0 suivantes tendront vers l'infini en n, et les $n-(\mathrm{r}_0+\mathrm{r}_1)$ plus petites valeurs propres seront bornées. Ainsi les facteurs (non stationnaires et stationnaires) et les *loadings* pourraient être estimés de façon standard par ACP. Dans la pratique, avec un petit nombre d'observations, et des facteurs non stationnaires qui chargeraient peu certaines séries par rapport à un facteur stationnaire qui les chargerait toutes beaucoup, on peut cependant tout à fait imaginer que l'ordre des valeurs propres ne correspondent au résultat attendu et que l'ensemble des premiers facteurs ne corresponde pas aux facteurs non stationnaires que l'on souhaite retenir.

Comme nous l'avons indiqué au paragraphe précédent, cette question est aussi abordée dans Banerjee et Marcellino (2009) ou dans Banerjee et *al.* (2010) qui proposent, eux, d'estimer d'abord les facteurs I(1), puis d'estimer des facteurs stationnaires sur les résidus de la régression des variables I(1) sur les facteurs I(1). Ces facteurs stationnaires sont ensuite regroupés avec ceux qui sont par ailleurs obtenus sur une base qui contient les variables stationnaires de l'échantillon, ainsi que les différences premières des variables non stationnaires. Cette méthode semble assez pertinente, mais d'une part on peut se demander s'il est effectivement souhaitable d'ajouter aux variables stationnaires les différences premières des variables non stationnaires, et d'autre part il ne semble pas forcément justifié d'estimer séparément les facteurs stationnaires associés aux résidus et les facteurs associés aux variables stationnaires : on pourrait aussi estimer les facteurs communs associés à un ensemble de variables comprenant à la fois ces résidus et les variables stationnaires de l'ensemble de données initiales.

2. Le cadre non stationnaire génère un certain nombre de difficultés

2.1 Un grand nombre de variables macroéconomiques usuelles ne sont pas stationnaires

La base de données retenue dans cette étude correspond à celle de l'étude de Combes, Doz et Fournier (2012) proche de celle qui avait été élaborée pour l'étude de Bessec et Doz (2011).

La base est donc constituée d'une centaine de variables comprenant :

- des soldes d'enquêtes : les principaux soldes des enquêtes de l'Insee rentrant dans la construction des indicateurs synthétiques dans l'industrie, les services, le bâtiment, le commerce de détail et l'enquête auprès des ménages ; mais également des soldes de l'enquête de la Banque de France dans l'industrie et les services et des transformations de tous ces soldes d'enquêtes ;
- des variables réelles : la consommation des ménages en produits manufacturés et ses composantes, les immatriculations de véhicules neufs, l'indice de production industrielle et ses composantes et des variables relatives au marché du travail ;
- des variables nominales monétaires et financières : les taux d'intérêt, la pente des taux, plusieurs indices boursiers, un indice de volatilité du marché, les agrégats monétaires et des indices de prix ;
- des indicateurs de l'environnement international : les taux de change de l'euro par rapport aux grandes devises et des indicateurs sur l'économie allemande et américaine.

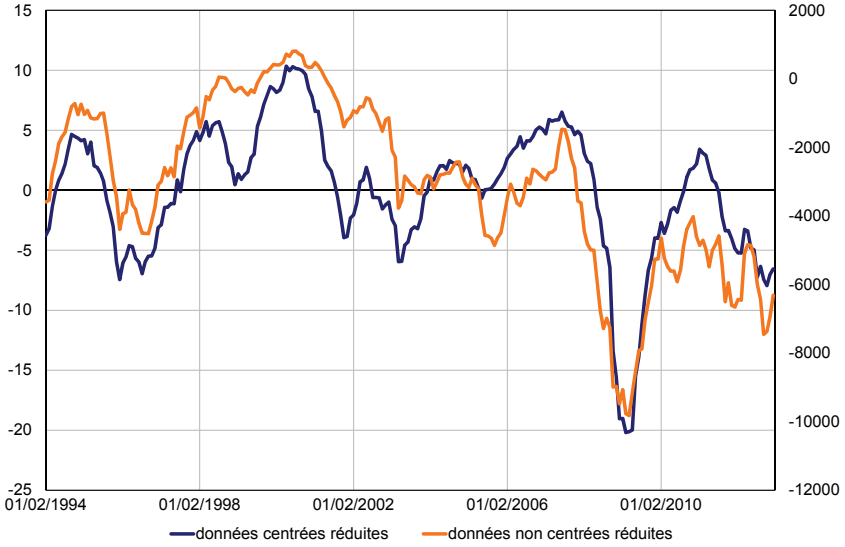
Les estimations présentées ici sont réalisées sur les séries telles qu'elles sont publiées aujourd'hui. Faute de séries disponibles, nous n'étudions pas l'impact des révisions des données sur les résultats. Les variables sont corrigées des variations saisonnières et pour la plupart sont de fréquence mensuelle. Certaines variables financières publiées à une fréquence journalière ou hebdomadaire sont mensualisées en prenant la dernière observation du mois. À noter enfin que les délais de publication diffèrent selon les séries : certaines variables sont disponibles au cours du mois qu'elles renseignent (les soldes d'enquêtes et les variables financières) mais les variables réelles sont connues avec un ou deux mois de retard.

Une analyse de la stationnarité des variables a été réalisée. Nous avons appliqué plusieurs tests de racine unitaire (tests ADF, Phillips Perron et KPSS) à l'ensemble des séries (prises en logarithme pour la plupart). Afin de contrôler l'effet de la dernière récession conduisant en fin de période à un saut en niveau sévère dans la plupart des séries, les tests ont été réalisés sur deux périodes : 1980-2010 ou 1980-2007. Les conclusions sont résumées dans l'annexe 1. À peu d'exceptions près, les soldes d'enquête sont considérés comme stationnaires, les variables réelles et financières intégrées d'ordre un (sauf la pente des taux d'intérêt qui est stationnaire selon la plupart des tests).

2.2 Des méthodes d'estimation non robustes aux changements d'échelle

La présence de variables non stationnaires aux moments divergents pose problème dans cette étude dans la mesure où les résultats de l'ACP traditionnellement utilisée dans un cadre stationnaire ne sont pas robustes aux changements d'échelle. À titre d'illustration, si l'on effectue une ACP sur les données essentiellement stationnaires de la base décrite précédemment, on constate que centrer réduire ne change pas beaucoup les résultats pour le premier facteur comme on peut le voir sur le graphique ci-dessous. Il n'en est pas de même sur une base de variables non stationnaires.

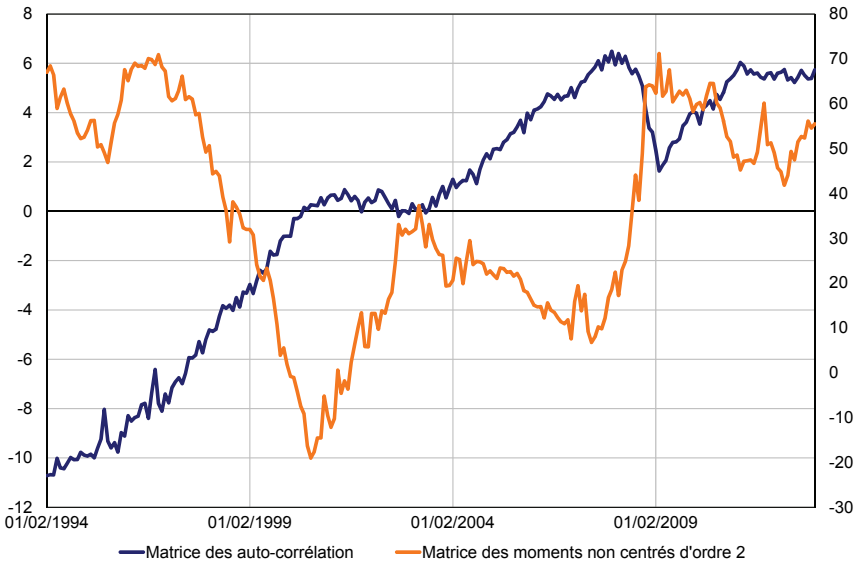
Graphique 1 : premiers facteurs calculés sur données stationnaires centrées réduites ou non



Source : DG Trésor.

S'il ne paraît pas raisonnable de centrer ou de réduire des variables aux moments divergents, il est gênant d'utiliser les facteurs calculés à partir de la matrice des moments non centrés d'ordre deux car leur calcul varie avec l'unité de mesure retenue pour les variables. On constate notamment que les résultats sont très différents de ceux obtenus par ACP sur données centrées-réduites.

Graphique 2 : premiers facteurs calculés des données I(1) centrées réduites ou non

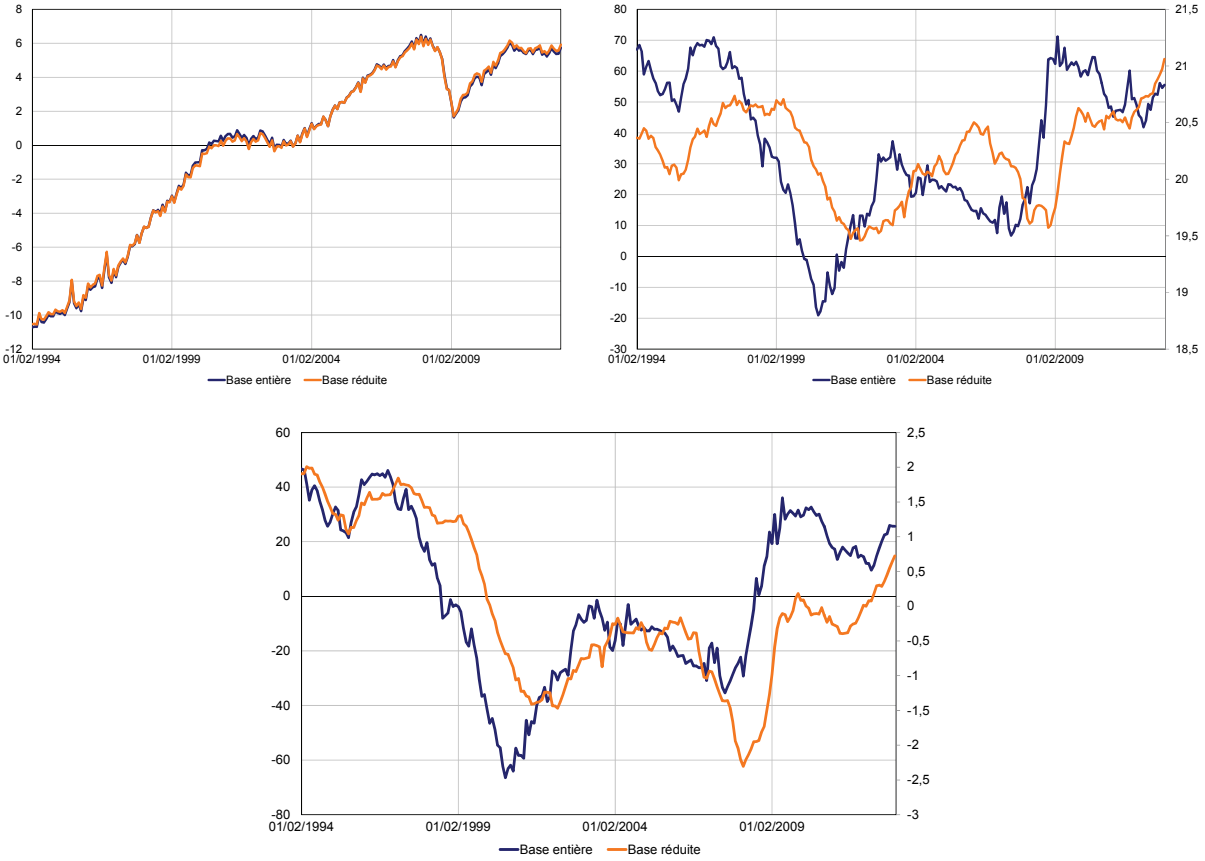


Source : DG Trésor.

En particulier, les résultats du calcul des facteurs à partir de la matrice des moments non centrés d'ordre deux diffèrent sensiblement si on modifie légèrement la base des variables non stationnaires. Ci-dessous, on a représenté les résultats pour une base avec ou sans les trois variables caractérisées par la plus grande volatilité[7].

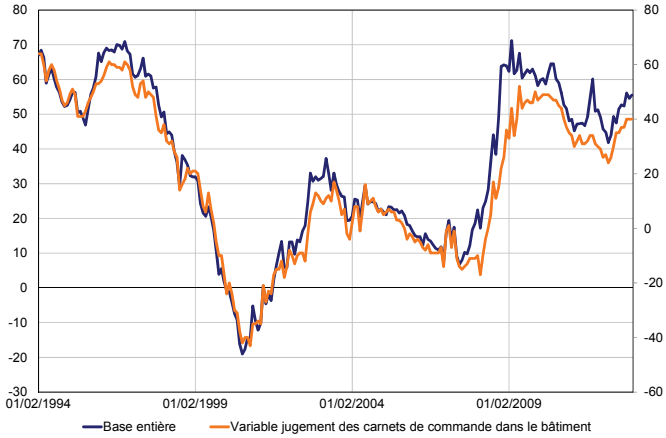
[7] Les résultats de comparaison de méthodes sur des cas théoriques simples sont présentés en annexe 2.

Graphiques 3 : calcul des premiers facteurs sur différentes bases par ACP sur matrice d'autocorrélations empiriques (en haut à gauche), à partir de la matrice des moments non centrés d'ordre 2 (en haut à droite), à partir de la matrice des covariances empiriques (en bas)



Source : DG Trésor.

Graphique 4 : les facteurs estimés sur la base entière sont proches des variables les plus volatiles



Source : DG Trésor.

2.3 Le traitement conjoint ou disjoint de variables stationnaires et non stationnaires n'est pas tranché dans la littérature

Comme on l'a vu dans les parties 1.2.3 et 1.2.4, il y a principalement deux méthodes proposées dans la littérature pour estimer les facteurs sur données I(1) :

- la méthode proposée par Bai (2004) qui consiste à extraire les facteurs communs à partir de la matrice des moments non centrés d'ordre deux calculée sur une base contenant les variables non stationnaires prises en niveau, sous l'hypothèse que les composantes idiosyncratiques soient stationnaires ;

- celle de Bai et Ng (2004) qui proposent au contraire de réaliser l'ACP sur des transformations stationnaires des variables non stationnaires et de cumuler ensuite les facteurs obtenus par ACP sur cette base. Cette approche permet, comme expliqué précédemment, de ne pas faire l'hypothèse de stationnarité des composantes idiosyncratiques.

Banerjee (2009) propose également une variante de la méthode proposée par Bai (2004), qui consiste à cumuler les variables stationnaires avant de les adjoindre à la base des variables non stationnaires, pour appliquer la même approche à un ensemble de variables non stationnaires.

Décrites sur le plan théorique en 1.2.3 et 1.2.4, on peut illustrer l'application de ces différentes méthodes sur données réelles et l'on constate qu'elles fournissent des résultats sensiblement différents (*cf.* Annexe 2).

De plus, ces méthodes apportent chacune une réponse différente à la question de traiter les variables ensemble ou séparément. Certaines isolent les variables non stationnaires tandis que d'autres les intègrent avant ou après les avoir cumulées.

Outre que les arguments en vertu d'une approche ou d'une autre sont rarement énoncés, notons par ailleurs, que la deuxième méthode semble, en particulier, peu convaincante puisqu'elle repose sur l'utilisation de facteurs calculés sur une base de variables différenciées pour reconstruire des facteurs non stationnaires, alors même que l'intuition qui prévaut pour privilégier l'approche en niveau provient de ce que l'on perd de l'information à différencier systématiquement les variables non stationnaires.

D'une manière générale, si l'approche par ACP se justifie bien dans le cadre stationnaire, on voit donc bien qu'elle pose un certain nombre de problèmes dans le cadre non stationnaire. Pour illustrer ceci, des simulations présentées en annexe 5 ont été réalisées et on peut voir que les résultats sont très sensibles aux paramètres retenus pour la simulation.

2.4 Le filtre de Kalman adapté

Notons que les aspects abordés précédemment concernent pour l'essentiel la détermination de la valeur initiale de l'estimation des facteurs, puisque dans la procédure en deux étapes, l'étape initiale détermine les valeurs des paramètres du modèle. Cette étape est néanmoins cruciale puisqu'elle détermine un certain nombre de paramètres.

Le filtre de Kalman peut être utilisé dans le cadre non stationnaire : la seule modification à apporter consiste à modifier la matrice de variance-covariance initiale du vecteur d'état (ici le vecteur des facteurs). En effet le filtre calcule de façon récursive, pour chaque date, la meilleure approximation du vecteur d'état et de la matrice de variance covariance de l'écart entre cette approximation et la vraie valeur (matrices P_t et $P_{t|t-1}$ de l'annexe 7). Il faut donc donner une valeur initiale à ces matrices (c'est-à-dire fixer la valeur de P_1 ou de $P_{1|0}$).

Dans le cas stationnaire, si l'on note f le vecteur d'état, on initialise en prenant pour P_1 la valeur commune des matrices de variance-covariance des f_t : cette valeur se calcule facilement en fonction des paramètres du modèle qui a été fixé pour les f_t.

Dans le cas non-stationnaire, on sait que la matrice de variance-covariance des f_t n'est pas constante, et qu'elle tend vers l'infini avec *t*. Par conséquent, on est amené à considérer ce qu'on appelle un *a priori* diffus (*diffuse prior*), c'est-à-dire à considérer une matrice P_1 infinie. En pratique, si f_t est un vecteur de dimension *q* modélisé sous la forme d'une marche aléatoire, on pose $P_1 = Vf_1 = \kappa I_q$, ou κ est une constante qui tend vers l'infini (il suffit de donner à κ une valeur numérique très grande). Du fait des inversions de matrices présentes dans les formules du filtre, l'impact de cette constante κ disparaît au bout de quelques itérations (on peut se référer à Harvey (1991) ou à Durbin et Koopman (2001) pour une présentation détaillée de cette question). Dans le cadre de la présente étude, nous postulons que les facteurs admettent une représentation autorégressive de type VARIMA (1,1,0), ce qui impose de choisir un vecteur d'état de dimension plus grande et d'adapter la forme de la matrice P_1, tout en conservant le même principe d'un *a priori* diffus (*cf.* annexe 7).

3 Protocole retenu et équations de prévision

3.1 Initialisation de l'estimation des facteurs

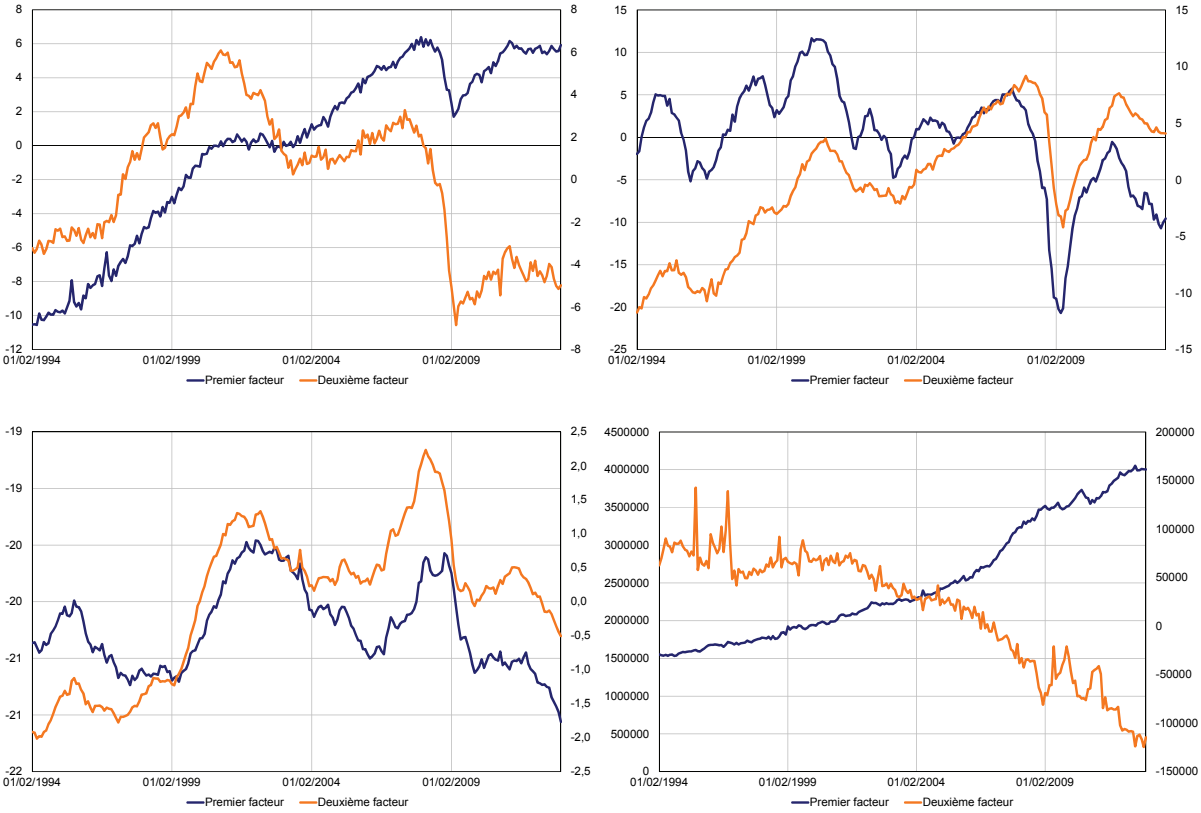
À défaut de pouvoir justifier théoriquement la meilleure méthode à mettre en œuvre pour traiter les données dans le cadre non stationnaire, différentes techniques d'estimation sont ici étudiées et comparées.

3.1.1 Les variables stationnaires et non stationnaires sont traitées séparément

Plusieurs raisons semblent indiquer qu'il est plus pertinent de traiter les variables stationnaires et non stationnaires séparément.

- Les différentes méthodes proposées pour l'extraction des tendances communes sont sensibles aux changements d'échelle. Ainsi l'ACP sur matrice de variance-covariance, ou le calcul des facteurs à partir de la matrice des moments non centrés d'ordre deux, semblent avoir tendance à écraser les variables stationnaires tandis que l'ACP sur matrice d'auto-corrélation leur donne trop d'importance.

Graphiques 5 : ACP sur données centrées réduites non stationnaires seulement (en haut à gauche), sur données centrées réduites stationnaires et non stationnaires (en haut à droite), calcul des facteurs à partir de la matrice des moments d'ordre 2 non centrés sur variables non stationnaires seulement (en bas à gauche), sur variables stationnaires et non stationnaires (en bas à droite)

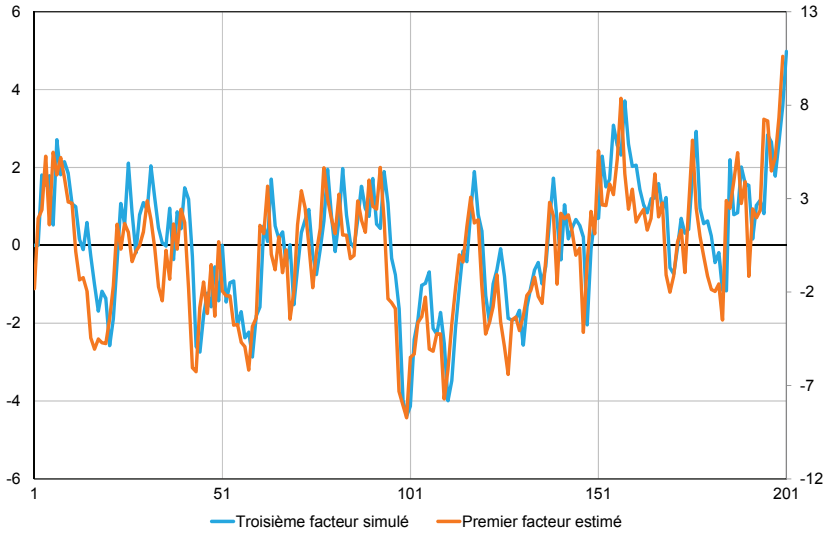


Source : DG Trésor.

- En outre, la modélisation retenue pour la prévision, le FECM, sépare les ajustements de court terme impliquant les facteurs stationnaires de l'équation de long terme qui rend compte de la cointégration entre la variable à prévoir et les facteurs non stationnaires. Dans ce cadre, il ne semble pas pertinent d'estimer des facteurs qui intègreraient des composantes communes à la fois non stationnaires et stationnaires.

On voit notamment sur les simulations réalisées en annexe 5 que le facteur stationnaire commun est mieux approché si l'on applique les techniques d'estimation sur la base des seules variables stationnaires.

Graphique 6 : facteurs estimés par ACP sur la matrice d'autocorrélation des seules variables stationnaires



Source : DG Trésor.

- Enfin, dans la littérature, il est souvent supposé que les premiers facteurs correspondent nécessairement à des facteurs non stationnaires, les facteurs stationnaires venant ensuite, partant du principe que les valeurs propres des facteurs non stationnaires chargeant un grand nombre de variables divergent à une vitesse plus importante que celles qui sont associées aux facteurs stationnaires (*cf.* 1.4). À distance finie (*i.e.* sur un nombre réduit d'observations), on peut toutefois imaginer que les valeurs propres ne soient pas rangées dans cet ordre, en particulier si l'on a un grand nombre de variables stationnaires dans la base ou si la variance d'un facteur stationnaire se trouve être supérieure à celle d'un facteur non stationnaire sur l'échantillon de données disponibles.

Pour ces raisons, il semble raisonnable de traiter séparément les variables non stationnaires et les variables stationnaires.

Nous avons ensuite retenu les techniques suivantes d'estimation des facteurs non stationnaires : l'extraction des facteurs communs à partir de la matrice des moments non centrés d'ordre 2 et l'extraction des facteurs communs à partir de la matrice des moments d'ordre 2 des variables prises en écart à leur valeur initiale (dans un souci de réduire les problèmes d'échelle).

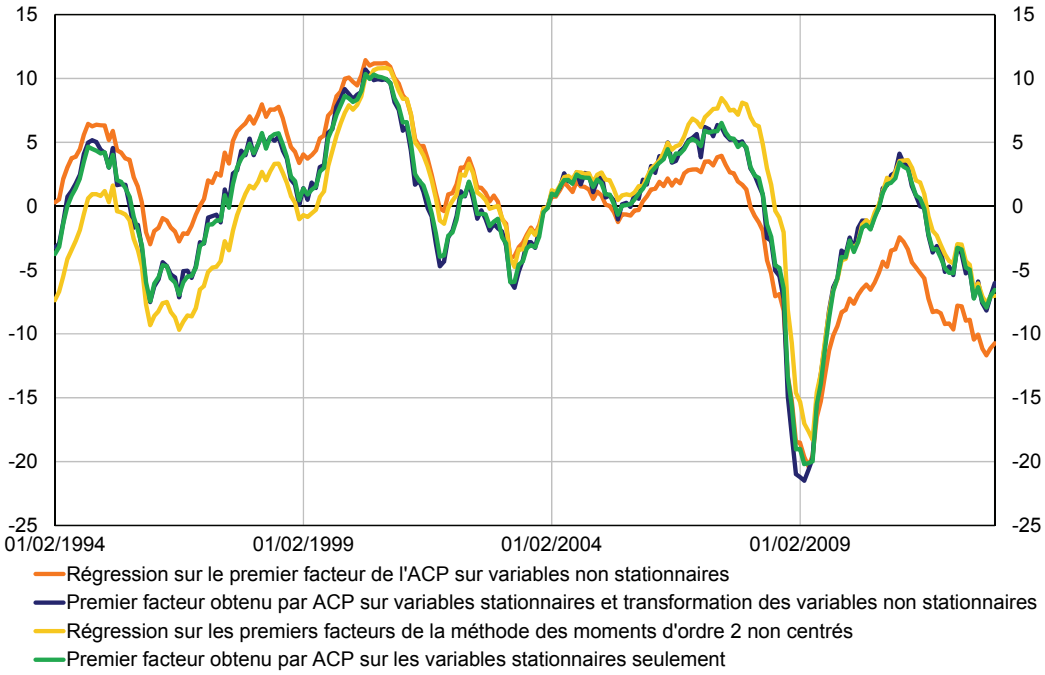
3.1.2 On adjoint à la base de variables stationnaires, les résidus de la régression des variables non stationnaires sur les facteurs I(1)

Pour les données stationnaires, la méthode est plus traditionnelle. Il s'agit tout de même de savoir comment on intègre la part stationnaire contenue dans les variables non stationnaires et qui peut être utile pour expliquer l'évolution du PIB à court terme. Une fois ce choix fait, on peut procéder comme dans le cadre stationnaire et calculer les facteurs stationnaires à partir d'une base comportant les variables stationnaires et les transformations stationnaires des variables non stationnaires.

Comme on peut imaginer que l'on perde une partie non négligeable de l'information en différenciant simplement les variables non stationnaires, on propose de les régresser sur les facteurs non stationnaires communs que l'on aura retenus en première étape. On pourra alors adjoindre les résidus issus de cette régression à la base des variables stationnaires, sous

l'hypothèse[8] que ceux-ci soient bien stationnaires. Le facteur stationnaire ainsi obtenu à partir des facteurs non stationnaires estimés en 3.1.1 est illustré ci-dessous (*cf.* annexe 4) :

Graphique 7 : comparaison du premier facteur de l'ACP sur variables stationnaires au premier facteur de l'ACP sur variables stationnaires et résidus de la régression des variables non stationnaires sur le premier facteur non stationnaire



Source : DG Trésor.

On peut, en calculant les contributions des variables au facteur, constater que dans les deux cas, les différences premières des variables non stationnaires ou les résidus des régressions interviennent bien dans le calcul des facteurs et qu'on a donc intérêt à prendre en compte ces variables même dans le cadre stationnaire.

3.2 Équations de prévision

3.2.1 Horizons de prévision

Dans le cadre de cette étude, comme dans Bessec-Doz (2011) et Combes-Doz-Fournier (2013), on s'intéresse à la prévision de l'évolution trimestrielle du PIB. Chaque horizon de prévision est défini par le nombre de mois à attendre avant l'obtention des premiers résultats du PIB publié par l'Insee environ 45 jours après la fin du trimestre concerné.

- Les prévisions pour le trimestre suivant (*Forecasting*) sont les prévisions du taux de croissance du trimestre *T* construites à partir des facteurs mensuels du trimestre *T-1*. On définit ainsi l'horizon H7 comme celui correspondant à la prévision faite à l'issue du mois 1 du trimestre *T-1*, l'horizon H6 à celle du mois 2 de *T-1* et H5 à celle du mois 3 de *T-1* ;
- Les prévisions du trimestre en cours (*Nowcasting*) consistent à prévoir la croissance du trimestre *T* à la fin de chaque mois de ce trimestre (H4 fin du mois 1, H3 fin du mois 2 et H2 fin du mois 3) ;
- Les prévisions pour le trimestre précédent (*Backcasting*) visent à estimer au cours du trimestre *T+1* la croissance du PIB du trimestre *T,* avant sa 1ère publication. Dans le cas qui nous occupe ici, cette prévision peut être réalisée à l'issue du 1er mois du trimestre *T+1*.

[8] Cette hypothèse est cruciale pour la validité de la régression et demande à être vérifiée empiriquement.

Tant que les premiers résultats du PIB ne sont pas publiés, les « prévisions » en *Backcasting* ont un intérêt pour le conjoncturiste (*cf.* schéma 1).

Schéma 1 : horizon des MFD dans le cas de la prévision du taux de croissance du PIB du trimestre T

<table>
<tr><td colspan="3">*FORECASTING* / trimestre T</td><td colspan="3">*NOWCASTING* / trimestre T</td><td colspan="2">*BACKASTING* / trimestre T</td></tr>
<tr><td colspan="3">Données du Trimestre T-1</td><td colspan="3">Données du Trimestre T</td><td colspan="2">Données du Trimestre T+1</td></tr>
<tr><td>Mois 1 / T-1</td><td>mois 2 / T-1 publications des Premiers résultats de T-2</td><td>Mois 3 / T-1</td><td>Mois 1 / T</td><td>mois 2 publications des Premiers résultats de T-1</td><td>Mois 3 / T</td><td>mois 1 / T+1</td><td>mois 2 / T+1</td></tr>
<tr><td>H7</td><td></td><td></td><td></td><td></td><td></td><td></td><td rowspan="7"></td></tr>
<tr><td></td><td>H6</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td>H5</td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td>H4</td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td>H3</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td>H2</td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td>H1</td></tr>
</table>

Source : DG Trésor.

3.2.2 Les équations de prévision avec ou sans prolongement des facteurs dans une approche en deux étapes

Dans l'approche en deux étapes que nous allons décrire ci-dessous, les équations sont proches du cas stationnaire déjà présenté dans le document de travail Combes-Doz-Fournier (2013), puisqu'il s'agit simplement d'adjoindre à l'équation de court terme le résidu de la régression de la variable d'intérêt en niveau sur les facteurs non stationnaires.

Si on note y_t la variable à prévoir (ici le PIB en niveau), on s'intéresse à la prévision faite pour Δy_{T+h} à la date *T*, avec $h = -1$ (*Backasting*), $h = 0$ (*Nowcasting*) ou $h = 1$ (*Forecasting*). Cette prévision est ici obtenue par l'intermédiaire d'une équation dite de « court terme » reliant les valeurs de la variable à prévoir à celles des facteurs stationnaires et du résidu de l'équation dite de « long terme » où l'on a régressé y_t sur les facteurs non stationnaires pris en niveau et trimestrialisés[9] (indicé *1*). On note η_t le résidu, et on vérifie qu'il est bien stationnaire par les tests usuels.

$$\hat{\eta}_t = y_t - \sum_{i=1}^{r_1} \hat{\gamma}_i f_{i,t}^{1,Q}$$

Les valeurs trimestrialisées des facteurs non stationnaires correspondent à la trimestrialisation des valeurs disponibles ou à la trimestrialisation des valeurs prolongées en fonction de la méthode utilisée (*cf. infra*).

Deux types d'équations de prévision peuvent être ensuite utilisés, dans l'esprit des approches rencontrées dans la littérature pour le cas stationnaire et reprises par Bessec et Doz :

La première approche consiste à estimer par MCO le modèle :

$$\Delta y_{t+h} = \delta_0 + \sum_{i=1}^{r_0} \delta_i f_{i,t}^{0,Q} + \alpha \hat{\eta}_{t-1} + \epsilon_{t+h} \quad (3a)$$

et à calculer ensuite la prévision de y_{T+h} à la date *T* en utilisant la formule suivante :

$$\Delta \hat{y}_{T+h|T} = \hat{\delta}_0 + \sum_{i=1}^{r_0} \hat{\delta}_i f_{i,T}^{0,Q} + \hat{\alpha} \hat{\eta}_{T-1} \quad (3b)$$

[9] Lorsque, comme c'est le cas ici, la variable à prévoir est une variable trimestrielle alors que les données utilisées, et donc aussi les facteurs estimés, sont à valeurs mensuelles, la prévision repose donc d'abord sur une *trimestrialisation* de ces facteurs. La valeur *trimestrialisée* f_t^Q du facteur stationnaire ou non stationnaire à la date *t* est, dans le cadre de cette étude, calculée comme une moyenne arithmétique des valeurs estimées aux différents mois du trimestre, ou en utilisant les taux de croissance mensuels (*cf.* Bessec-Doz (2012) ou Combes, Doz, Fournier (2013).

Lorsqu'on utilise cette approche dans le cadre de cette étude, on le fait en pratique en utilisant deux équations :

- pour la prévision du trimestre suivant (*Forecasting*), resp. courant (*Nowcasting*), à partir des équations (3a) et (3b) avec $h = 1$, *resp* $h = 0$;
- pour $h = -1$ (*Backcasting*), on estime l'équation (3a) pour $h = 0$, et on calcule la prévision sous la forme :

$$\Delta\hat{y}_{T-1|T} = \hat{\delta}_0 + \sum_{i=1}^{r_0} \hat{\delta}_i \, f_{i,T-1}^{0,Q} + \hat{\alpha} \, \hat{\eta}_{T-2} \quad (3c)$$

Dans le cadre de cette première approche, aucune prévision des facteurs n'est donc effectuée et seules les valeurs disponibles sont utilisées pour la trimestrialisation.

Une autre approche est spécifiquement liée au cadre dynamique et utilise l'estimation de la dynamique des facteurs stationnaires. On estime d'abord par les moindres carrés ordinaires l'équation reliant le taux de croissance du PIB aux facteurs *trimestrialisés* qui lui sont contemporains :

$$\Delta y_t = \delta_0 + \sum_{i=1}^{r_0} \delta_i \, f_{i,t}^{0,Q} + \alpha \, \hat{\eta}_{t-1} + \epsilon_t \quad (4a)$$

Cette équation coïncide avec l'équation (3a) lorsque celle-ci est estimée avec $h = 0$.

On calcule ensuite une prévision de y_{T+h} à la date *T* à partir d'une prévision du facteur *trimestrialisé*. Si les facteurs vérifient un modèle de la forme : $f_t = \sum_{i=1}^{p} A_i \, f_{t-i} + \varepsilon_t$, on obtient une prévision mensuelle $f_{T+m|T}$ de f_{T+m} à la date *T* pour tous les mois *m* couvrant la fin du trimestre en cours (si la prévision est faite aux mois 1 ou 2 du trimestre *T*) et le trimestre suivant. Cette prévision s'obtient de façon récursive en utilisant les valeurs estimées des matrices A_i et des facteurs ; on peut donc ensuite trimestrialiser les prévisions mensuelles pour déterminer une prévision $f_{i,T+h|T}^{Q}$ du facteur *trimestrialisé*. On obtient alors une prévision de y_{T+h} à la date *T* en utilisant la formule suivante :

$$\Delta\hat{y}_{T+h|T} = \hat{\delta}_0 + \sum_{i=1}^{r_0} \hat{\delta}_i \, f_{i,T+h|T}^{0,Q} + \hat{\alpha} \, \hat{\eta}_{T-1} \quad (4b)$$

Dans cette approche, la même formule est employée quelle que soit la valeur de h $(h = 1, 0 \; ou \; -1)$.

Dans une troisième approche, on utilise les équations (3a) et (3b), mais en modifiant la démarche lorsque la date *T* à laquelle la prévision est faite correspond au 1er ou au 2e mois d'un trimestre : dans ce cas, l'équation (3a) est utilisée comme précédemment, mais à partir d'une prévision des facteurs sur les mois suivants du trimestre concerné, estimée par la représentation VAR des facteurs. La prévision est ensuite obtenue en appliquant l'équation (3b) au facteur *trimestrialisé* associé.

On peut donc résumer les trois méthodes employées ici de la façon suivante :

- **Méthode 1 : utilisation d'une seule équation de prévision après prolongation des facteurs mensuels** (et leur *trimestrialisation*). On estime une seule équation (4a), qui relie les valeurs contemporaines du taux de croissance du PIB et des facteurs *trimestrialisés*. Les prévisions *Forecasting*, *Nowcasting* et *Backcasting* sont calculées de façon identique (équation 4b) en utilisant les valeurs prévues ou estimées des facteurs.
- **Méthode 2 : une équation de prévision spécifique pour chaque horizon de prévision sans prolongement des facteurs mensuels**. La prévision *Forecasting* est alors établie à partir de l'équation (3a) associée à $h = 1$, tandis que les prévisions *Nowcasting* et *Backcasting,* sont obtenues à partir de l'équation (3a) associée à $h = 0$.
- **Méthode 3 : une équation de prévision spécifique pour chaque horizon de prévision mais en prolongeant les facteurs mensuels à l'horizon du trimestre en cours**. Il s'agit d'une version mixte des deux méthodes précédentes donnant des

résultats différents seulement pour les 2 premiers mois du trimestre. Elle est identique à la méthode 2 dans le cas du 3e mois du trimestre.

On notera donc que pour le *Nowcasting*, la méthode 2 est identique à la méthode 1 dans le cas du mois 3 du trimestre, date à laquelle il n'y a pas besoin de prolonger les facteurs mensuels avant de les *trimestrialiser*; et que la méthode 3 n'a pas lieu d'être. Pour le *Backasting*, toutes les méthodes sont équivalentes puisque cette prévision est calculée pendant le mois qui suit le trimestre à prévoir.

Le schéma 3 (ci-dessous) résume les développements précédents et présente entre les différentes méthodes d'estimation disponibles en fonction de l'horizon de prévision visé.

Schéma 3 : méthodes d'estimation des modèles de prévision de la croissance du PIB du trimestre T

Données disponibles à la fin		Horizons de prévision		Méthode d'estimation des facteurs mensuels		
T-1	*mois 1*	*H7*	"Forecasting"	Méthode 1	Méthode 2	Méthode 3
T-1	*mois 2*	*H6*		Méthode 1	Méthode 2	Méthode 3
T-1	*mois 1*	*H5*		Méthode 1	Méthode 2 = Méthode 3	
T	*mois 1*	*H4*	"Nowcasting"	Méthode 1	Méthode 2	
T	*mois 2*	*H3*		Méthode 1	Méthode 2	
T	*mois 3*	*H2*		Méthode 1 = Méthode 2		
T+1	*mois 1*	*H1*	"Backcasting"	Méthode 1 = Méthode 2		

Source : DG Trésor.
Lecture :

- *Méthode 1 : une seule équation et les facteurs mensuels sont prolongés avant trimestrialisation ;*
- *Méthode 2 : deux équations, et les facteurs ne sont pas prolongés ;*
- *Méthode 3 : deux équations et les facteurs sont prolongés seulement sur le trimestre en cours.*

3.2.3 Les équations de prévision avec ou sans prolongement des facteurs dans l'approche FECM

Si l'on souhaite obtenir une estimation plus rigoureuse et plus robuste, on estime un modèle vectoriel à correction d'erreur (noté FECM lorsque le vecteur concerné contient des facteurs). Le nombre de facteurs non stationnaires et stationnaires sollicités dans cette démarche ainsi que leurs retards peuvent nous conduire à estimer un trop grand nombre de coefficients.

Cependant, sous hypothèse d'exogénéité faible des facteurs stationnaires (*cf.* annexe 6), on peut ne retenir que la première équation du modèle à correction d'erreur estimé par la méthode de Johansen.

Supposons qu'on a un vecteur $Z_t = \begin{pmatrix} z_t \\ f_t^0 \end{pmatrix}$ avec $z_t \sim I(1)$ cointégré et $f_t^0 \sim I(0)$, avec ici $z_t = \begin{pmatrix} y_t \\ f_t^1 \end{pmatrix}$ et f_t^1 et f_t^0 les facteurs non stationnaires et stationnaires respectivement. En toute rigueur, le modèle VECM associé peut s'écrire sous la forme suivante :

$$\begin{pmatrix} \Delta z_t \\ \Delta f_t^0 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} - \begin{pmatrix} \alpha_{11}\ \alpha_{12} \\ \alpha_{21}\ \alpha_{22} \end{pmatrix} \begin{pmatrix} \beta' & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} z_{t-1} \\ f_{t-1}^0 \end{pmatrix} + \sum_{k=1}^{p-1} \begin{pmatrix} c_{11}^k\ c_{12}^k \\ c_{21}^k\ c_{22}^k \end{pmatrix} \begin{pmatrix} \Delta z_{t-k} \\ \Delta f_{t-k}^0 \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix}$$

Si l'on ne retient que la première équation, les équations de prévision (4a) et (3a) utilisées pour les différentes méthodes décrites précédemment prennent la forme suivante :

$$\Delta y_{t+h} = A_1 \Delta f_t^0 + \tilde{\mu}_{1,1} + \tilde{\alpha}_{11,1}\beta' z_{t-1} + \sum_{k=1}^{p-1} \tilde{c}_{11,1}^k \Delta z_{t-k} + \sum_{k=1}^{p} \tilde{d}_{12,1}^k f_{t-k}^0 + \tilde{\epsilon}_{t+h} \quad \text{(3a bis)}$$

Et :

$$\Delta y_{t+h} = A_1 \Delta f_{t+h}^0 + \tilde{\mu}_{1,1} + \tilde{\alpha}_{11,1}\beta' z_{t-1} + \sum_{k=1}^{p-1} \tilde{c}_{11,1}^k \Delta z_{t-k} + \sum_{k=1}^{p} \tilde{d}_{12,1}^k f_{t+h-k}^0 + \tilde{\epsilon}_{t+h} \quad \text{(4a bis)}$$

3.3 Tests et évaluation des performances en pseudo-temps réel

Pour évaluer les performances d'une technique de prévision, il est classique de calculer les prévisions qui auraient été fournies par le modèle avec les données qui étaient effectivement disponibles à une date donnée. Ce calcul est fait ici sur une plage de dates couvrant dix années et les prévisions obtenues sont ensuite comparées avec les vraies valeurs du taux de croissance du PIB aux dates concernées.

Le pseudo-temps réel[10] permet également de vérifier la robustesse d'une modélisation.

De façon plus précise, l'analyse est réalisée comme suit. L'échantillon utilisé pour estimer le modèle à facteurs contient une centaine de variables conjoncturelles publiées à fréquence mensuelle sur la période 1990T1-2013T4. On se concentre d'abord sur l'échantillon des données correspondant à la période 1990T1-1999T4[11]. Les tests (ou l'estimation du modèle à facteurs) sont réalisés (ou estimés) sur l'ensemble de données ainsi constitué. Puis l'échantillon est augmenté pas à pas, d'un certain nombre d'observations à chaque étape, et pour chacun des sous-échantillons obtenus, tous les calculs sont répliqués : tests ou estimation du modèle à facteurs, calcul des prévisions et des erreurs de prévision. L'exercice est poursuivi jusqu'à ce que l'ensemble de l'échantillon disponible ait été utilisé. L'utilisation du filtre de Kalman permet de prendre en compte toutes les données disponibles à chaque date, en tenant compte de leurs délais de publication différents.

Ici, nous avons réalisé des tests de cointégration itératifs sur le PIB en niveau et un à quatre facteurs non stationnaires extraits des échantillons itératifs. Nous avons également testé l'exogénéité faible d'un vecteur comprenant un à quatre facteurs stationnaires et nous avons contrôlé la significativité et le signe de la force de rappel de la première relation de cointégration dans le VECM estimé dans le cas où l'hypothèse n'était pas rejetée. Ces tests itératifs ont été réalisés pour les quatre protocoles décrits en 3.1.2

Il en a résulté que la seule combinaison satisfaisante et offrant des résultats relativement stables dans le temps consistait à considérer deux facteurs non stationnaires et un facteur stationnaire.

En partie 4, nous avons renseigné les résultats en performances obtenus pour les différentes approches d'initialisation et pour les différentes modélisations (en deux étapes ou vectorielle).

[10] Il s'agit d'un exercice dit de « pseudo temps réel » car si l'on tient effectivement compte des délais de publications des séries, il n'est cependant pas possible de tenir compte des révisions apportées aux variables conjoncturelles au gré de leurs publications. Ces révisions peuvent être importantes dans le temps, notamment pour les variables ayant le plus d'influence sur la variable cible (IPI par exemple).

[11] Par simplicité on considère pour les tests un échantillon cylindré, c'est-à-dire que l'on considère que toutes les données sont disponibles jusqu'à la fin du 1999T4. Or on sait que dans les faits les délais de publication varient en fonction des variables. Cet aspect est pris en compte pour la mesure des indicateurs de performance des modèles (RMSFE) *cf.* partie 4.

4 Impact sur les performances en prévision du choix des blocs de variables utilisés

Dans le document de travail Combes, Doz, Fournier (2013), on avait pu mettre en évidence le fait que la sélection de données, en fonction de l'horizon de prévision et préalable à l'utilisation des MFD, améliore les performances en prévision par rapport aux seules trois combinaisons de blocs de variables retenues dans l'étude de 2011.

Nous avons donc conservé dans cette étude les huit blocs définis dans l'étude précédente, et comme dans celle-ci, nous avons comparé les performances des modèles utilisant toutes les combinaisons possibles de huit blocs de variables stationnaires, soit 255 combinaisons constituées de un à huit blocs de variables, à chaque horizon de prévision.

Pour rappel, les huit blocs de variables mensuelles considérés dans l'étude précédente étaient les suivants (*cf.* annexe 1 de Doz-Combes-Fournier pour les blocs 1 à 6 et annexe 3 pour les blocs 7 et 8) :

- **Bloc 1** (**23 variables mensuelles, disponibles depuis janvier 1991)** : *constitué des principaux soldes d'opinion issus des enquêtes mensuelles de conjoncture de l'Insee*.
- **Bloc 2** (**25 variables mensuelles disponibles depuis décembre 1990)** : *constitué des variables réelles, dont la plupart sont utilisées en tant qu'inputs par les comptables nationaux pour construire les comptes trimestriels et déterminer le taux de croissance trimestriel du PIB.*
- **Bloc 3** (**22 variables mensuelles disponibles depuis janvier 1991)** : *constitué de variables nominales* monétaires et financières (taux d'intérêt, pente des taux, indices boursiers, etc.).
- **Bloc 4** (**16 variables sont prises en compte dans ce bloc depuis janvier 1992)** : *indicateurs de l'environnement international* (taux de change de l'euro et indicateurs conjoncturels des principaux partenaires économiques).
- **Bloc 5** (**20 variables, disponibles depuis février 1991)** : *variations mensuelles des soldes d'opinion des enquêtes de conjoncture de l'Insee*.
- **Bloc 6** (**20 variables, disponibles depuis janvier 1991)** : *soldes d'opinion des enquêtes de conjoncture de l'Insee au carré signé (à savoir les soldes pris au carré mais dont le signe est conservé*) : ceci permet d'introduire dans les modèles des éléments de non-linéarité[12].
- **Bloc 7** (**12 variables disponibles depuis septembre 1991)** : *soldes d'opinion et indices de climat des affaires issus des enquêtes de conjoncture réalisées par la Banque de France auprès des chefs d'entreprises*.
- **Bloc 8** (**11 variables disponibles depuis octobre 1991)** : *évolution mensuelle des soldes d'opinion et des indices de climat des affaires issus des enquêtes de conjoncture réalisées par la Banque de France auprès des chefs d'entreprises*. Comme pour les résultats des enquêtes Insee, ces variables correspondent aux différences premières des soldes d'opinion mensuels.

Les facteurs non stationnaires sont extraits, quant à eux, de la base constituée de l'ensemble des variables non stationnaires. Ces dernières n'ont pas été réparties en blocs distincts, car elles sont moins nombreuses.

[12] Ce type d'approche avait été retenue par Bai and Ng (2008), "Forecasting Economic Time Series Using Targeted Predictors", *Journal of Econometrics*, 146.

4.1 Performance en prévision des modèles dans le cadre stationnaire

Afin de mesurer les gains potentiels en termes de performance en prévision de l'approche dans le cadre non stationnaire, nous avons actualisé le modèle habituellement utilisé en prévision par le bureau Prev3 avec les comptes annuels 2012 parus au mois de mai dernier. Le paramétrage de ce modèle (combinaison, méthode...) constitue donc un nouveau benchmark.

L'indicateur de performance en prévision utilisé ici est le RMSFE qui mesure l'erreur de prévision moyenne en pseudo-temps réel (définition rappelée en 3.1.1).

4.1.1 Résultats obtenus avec les combinaisons de variables optimales

Les résultats ci-dessous constituent un nouveau *benchmark* permettant de comparer les gains potentiels en performance des modèles tenant compte de la non-stationnarité des variables utilisées (*cf.* 3.2).

Tableau 1 : modèles les plus précis au sens du RMFSE pour le nouveau *benchmark*

	Horizon	RMFSE	Combi-naison	Méthode	Variables
Forecasting (m1 / T-1)	**H7**	0,35	220	M3	Enquêtes Insee et Banque de France, variables réelles, nominales et internationales
Forecasting (m2 / T-1)	**H6**	0,38	168	M3	Enquêtes Insee, variables réelles, nominales et internationales
Forecasting (m3 / T-1)	**H5**	0,36	187	M1	Enquêtes Insee et Banque de France, variables nominales et internationales
Nowcasting (m1 / T)	**H4**	0,29	253	M1	Enquêtes Insee et Banque de France, variables réelles, nominales et internationales
Nowcasting (m2 / T)	**H3**	0,26	113	M1	Enquêtes Insee et Banque de France, variables réelles
Nowcasting (m3 / T)	**H2**	0,23	142	M1	Enquêtes Insee et Banque de France, variables réelles et nominales
Backcasting (m1 / T+1)	**H1**	0,21	142	M1	Enquêtes Insee et Banque de France, variables réelles et nominales

Lecture des tableaux :
M1 : une seule équation et les facteurs mensuels sont prolongés avant trimestrialisation ;
M2 : deux équations pour les trimestres courant et suivant, facteurs non prolongés ;
M3 : deux équations pour les trimestres courant et suivant avec prolongement des facteurs (trimestre en cours).
Source : DG Trésor.

4.1.2 Résultats obtenus sur une base de données stationnaires alternatives

Ces résultats s'appuient sur une base de données constituée des variables stationnaires et des résidus des variables non stationnaires régressées sur les facteurs non stationnaires à partir de facteurs communs des variables non stationnaires. Les blocs restent conformes à la définition donnée en 3.1, seule la transformation de la variable considérée, lorsqu'elle est non stationnaire, change.

Cette approche, plus simple de mise en œuvre que le FECM, permet de voir si la différenciation systématique des variables non stationnaires ne fait pas déjà perdre une partie importante de l'information utile pour la prévision.

Les prévisions sont de qualité comparable voire moins bonne, en particulier lorsque l'on se rapproche de la publication. La régression sur les facteurs non stationnaires, qui peuvent posséder une composante stationnaire, entraîne finalement une perte d'information.

Tableau 2 : modèles les plus précis au sens du RMFSE pour la base de variables stationnaires alternatives

	RMFSE
Forecasting (m1 / T-1)	0,40
Forecasting (m2 / T-1)	0,38
Forecasting (m3 / T-1)	0,39
Nowcasting (m1 / T)	0,30
Nowcasting (m2 / T)	0,31
Nowcasting (m3 / T)	0,30
Backcasting (m1 / T+1)	0,27

Source : DG Trésor.

4.2 Performance en prévision des modèles dans le cadre non stationnaire

4.2.1 Résultats obtenus avec les combinaisons optimales pour l'approche en deux étapes

L'utilisation des facteurs non stationnaires sous la forme d'un résidu d'équilibre de long terme, ne permet pas de remédier à la perte d'information, comme on peut le voir dans le tableau ci-dessous.

Tableau 3 : modèles les plus précis au sens du RMFSE

	Matrice des moments d'ordre deux non centrés		Matrice des moments d'ordre deux non centrés sur variables prises en écart par rapport à leur valeur initiale
	Horizon	RMFSE	RMSFE
Forecasting (m1 / T-1)	**H7**	0,40	0,42
Forecasting (m2 / T-1)	**H6**	0,39	0,37
Forecasting (m3 / T-1)	**H5**	0,39	0,37
Nowcasting (m1 / T)	**H4**	0,30	0,32
Nowcasting (m2 / T)	**H3**	0,32	0,32
Nowcasting (m3 / T)	**H2**	0,30	0,30
Backcasting (m1 / T+1)	**H1**	0,27	0,30

Source : DG Trésor.

On peut voir par ailleurs que les aménagements réalisés sur les données non stationnaires entraînent une relative robustesse du modèle aux différentes approches appliquées pour extraire les facteurs communs non stationnaires. Ainsi, les RMSFE sont relativement proches selon qu'on utilise les matrices des moments d'ordre deux non centrés ou les matrices des moments d'ordre 2 non centrés pris sur les variables en écart par rapport à leur valeur initiale.

4.2.2 Résultats obtenus avec les combinaisons optimales pour l'approche vectorielle

L'approche vectorielle ne permet malheureusement pas d'améliorer les performances, le nombre de paramètres à estimer doit faire perdre de la précision aux estimations, ainsi les modèles apparaissent légèrement moins performants que les précédents à quelques mois de la publication et beaucoup moins performants à plus long terme.

Tableau 4 : modèles les plus précis au sens du RMFSE

	Matrice des moments d'ordre deux non centrés		Matrice des moments d'ordre deux non centrés sur variables prises en écart par rapport à leur valeur initiale
	Horizon	RMFSE	RMSFE
Forecasting (m1 / T-1)	H7	0,47	0,48
Forecasting (m2 / T-1)	H6	0,46	0,43
Forecasting (m3 / T-1)	H5	0,44	0,43
Nowcasting (m1 / T)	H4	0,37	0,37
Nowcasting (m2 / T)	H3	0,33	0,32
Nowcasting (m3 / T)	H2	0,31	0,30
Backcasting (m1 / T+1)	H1	0,30	0,29

Source : DG Trésor.

Conclusion

Dans une précédente étude, avaient été examinées les performances en prévision des modèles à facteurs dynamiques pour la prévision du taux de croissance du PIB français sur des horizons courts, en utilisant une base constituée d'une centaine de variables comprenant des variables d'enquêtes, des indicateurs réels, des variables monétaires et financières et des indicateurs sur l'environnement international. L'ensemble de ces variables étaient stationnaires ou rendues stationnaires par une transformation. Or il est bien connu que la plupart des séries macroéconomiques doivent être considérées comme non stationnaires. Depuis l'introduction de la notion de cointégration, on sait que, lorsqu'on travaille avec des variables non stationnaires, il peut être utile de prendre en compte l'information sur les relations d'équilibre de long terme entre ces variables. Il pouvait donc sembler utile de combiner les avantages des MFD (la prise en compte d'un grand nombre de variables) et des modèles à correction d'erreur (la prise en compte des relations de long terme ou de cointégration) pour la prévision de court terme. Après avoir réalisé une analyse précise de la base de données et essayé de déterminer empiriquement le nombre de facteurs non stationnaires qui pourraient sous-tendre à l'évolution conjointe des variables non stationnaires, nous avons mis en œuvre un certain nombre d'aménagements aussi bien au niveau de l'initialisation du modèle, du filtre de Kalman et des équations de prévision. Il s'est avéré que les résultats obtenus étaient moins bons que les résultats qui avaient été obtenus précédemment en n'utilisant que les transformations stationnaires des variables étudiées. Il est possible que cette détérioration des résultats provienne de la plus grande complexité des modèles faisant intervenir les relations de long terme entre la variable à prévoir et les facteurs non stationnaires. Néanmoins, des pistes d'amélioration mériteraient d'être étudiées en complément, notamment en modifiant la spécification des équations de prévision. Tout d'abord, on pourrait utiliser des facteurs stationnaires en plus grand nombre, afin d'exploiter davantage d'information. Par ailleurs, on pourrait modifier le terme de correction d'erreur lorsqu'il s'agit de prévoir le trimestre en cours ou le trimestre suivant et que la valeur de la variable d'intérêt au trimestre précédent n'est pas encore connue : au lieu d'utiliser un terme de correction d'erreur décalé de deux trimestres, il est probablement possible d'utiliser un terme de correction d'erreur fondé sur la prévision de la variable au trimestre précédent.

Annexe 1 : Ordre d'intégration des variables utilisées

Type	Secteur	série	stationnaire	Type	Secteur	série	stationnaire
Enquête	Industrie	activité passée	X	réel	ventes détail	ventes de détail	I(1)
Enquête	Industrie	stocks	X	réel	chômage	chômage	I(1)
Enquête	Industrie	commandes	X	réel	chômage	chômage des jeunes	I(1)
Enquête	Industrie	carnets de commande étrangers	X	réel	emploi	emplois vacants	I(1)
Enquête	Industrie	perspectives personnelles	X	réel	Industrie	BE IPI	I(1)
Enquête	Industrie	perspectives générales	X	réel	Industrie	CZ IPI	I(1)
Enquête	Services	perspectives générales	X	réel	Industrie	C1 IPI	I(1)
Enquête	Services	activité passée	X	réel	Industrie	C2 IPI	I(1)
Enquête	Services	perspectives personnelles	X	réel	Industrie	C3 IPI	I(1)
Enquête	Services	demande prévue	X	réel	Industrie	CL1 IPI	I(1)
Enquête	Bâtiment	activité passée	X	réel	Industrie	CL2 IPI	I(1)
Enquête	Bâtiment	perspectives personnelles	X	réel	Industrie	C5 IPI	I(1)
Enquête	Bâtiment	effectifs passés	X	réel	Industrie	DE IPI	I(1)
Enquête	Bâtiment	jugement sur les carnets de commande	I(1)	réel	Industrie	F IPI	I(1)
Enquête	Bâtiment	taux d'utilisation des capacités	I(1)	réel	commerce extérieur	AZ export	I(1)
Enquête	commerce de détail	perspectives générales	X	réel	commerce extérieur	AZ import	I(1)
Enquête	commerce de détail	ventes passées	X	réel	commerce extérieur	C1 export	I(1)
Enquête	commerce de détail	commandes	X	réel	commerce extérieur	C1 import	I(1)
Enquête	commerce de détail	effectifs prévus	X	réel	commerce extérieur	C2 export	I(1)
Enquête	ménages	situation financière passée	X	réel	commerce extérieur	C2 import	I(1)
Enquête	ménages	situation financière prévue	X	réel	commerce extérieur	C3 export	I(1)
Enquête	ménages	niveau de vie passée	X	réel	commerce extérieur	C3 import	I(1)
Enquête	ménages	niveau de vie prévue	X	réel	commerce extérieur	C4 export	I(1)
Enquête	ménages	opportunité d'achat	I(1)	réel	commerce extérieur	C4 import	I(1)
réel	Immatriculations	immatriculations	I(1)	réel	commerce extérieur	C5 export	I(1)
réel	consommation manufacturière	consommation manufacturière	I(1)	réel	commerce extérieur	C5 import	I(1)
réel	consommation manufacturière	consommation de biens durables	I(1)	réel	commerce extérieur	DE export	I(1)
réel	consommation manufacturière	consommation automobile	I(1)	réel	commerce extérieur	DE import	I(1)
réel	consommation manufacturière	consommation équipement logement	I(1)	nominal	bourse	cac40	I(1)
réel	consommation manufacturière	consommation textiles	I(1)	nominal	bourse	SP500	I(1)
réel	consommation manufacturière	consommation autres produits manufacturés	I(1)	nominal	bourse	FTSE	I(1)

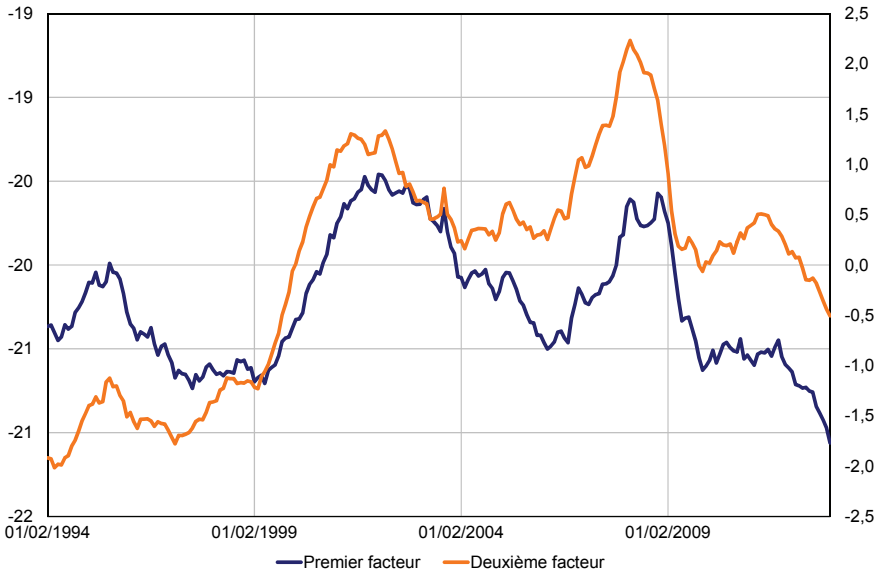
Type	Secteur	série	stationnaire	Type	Secteur	série	stationnaire
nominal	bourse	DAX	I(1)	International	taux de change	euro livre	I(1)
nominal	bourse	eurostoxx50	I(1)	International	taux de change	euro yen	I(1)
nominal	bourse	Nikkei	I(1)	International	taux de change	euro yuan	I(1)
nominal	bourse	PER_US	I(1)	International	taux de change	taux de change effectif	I(1)
nominal	bourse	vix	X	International	taux de change	taux de change effectif réel	I(1)
nominal	monnaie	M1	I(1)	International	Allemagne	activité passée IFO	X
nominal	monnaie	M2	I(1)	International	Allemagne	perspectives personnelles IFO	X
nominal	monnaie	M3	I(1)	International	Allemagne	activité passée ZEW	X
nominal	monnaie	prets	I(1)	International	Allemagne	perspectives personnelles ZEW	X
nominal	taux d'intérêt	Tlimmo	I(1)	International	Allemagne	confinace	X
nominal	taux d'intérêt	TI3M	I(1)	International	Allemagne	IPI manufacturier	I(1)
nominal	taux d'intérêt	TI1Y (gov. Bond)	I(1)	International	US	ventes de détail	I(1)
nominal	taux d'intérêt	TI10Y (gov. Bond)	I(1)	International	US	IPI manufacturier	I(1)
nominal	taux d'intérêt	spreadFR	X	International	US	Emploi	I(1)
nominal	taux d'intérêt	spreadUS	X	International	US	chômage	I(1)
nominal	prix	or	I(1)	International	US	PMI manufacturiers	X
nominal	prix	oil	I(1)	Réel	Ensemble	PIB France	I(1)
nominal	prix	raw	I(1)	Réel	Ensemble	Emploi	I(1)
nominal	prix	IPC	I(1)	Nominal	Ensemble	Salaires	I(1)
International	taux de change	euro dollar	I(1)				

Annexe 2 : Premiers facteurs obtenus dans les différents protocoles retenus pour réaliser l'ACP

Dans l'ensemble de la littérature se rapportant à ce sujet, nous avons relevé trois protocoles différents pour le traitement des facteurs non stationnaires. À noter que, dans notre modélisation, ces techniques n'ont vocation qu'à donner une valeur initiale à l'estimation des facteurs, puisqu'ensuite ils sont réestimés par filtre de Kalman. Cette étape est néanmoins cruciale puisqu'elle détermine les valeurs des paramètres du modèle.

- Cas n°1 : méthode proposée par Bai (2004) qui consiste à utiliser la matrice des moments non centrés d'ordre deux sur une base contenant les variables non stationnaires en niveau, sous l'hypothèse que les composantes idiosyncratiques sont stationnaires.

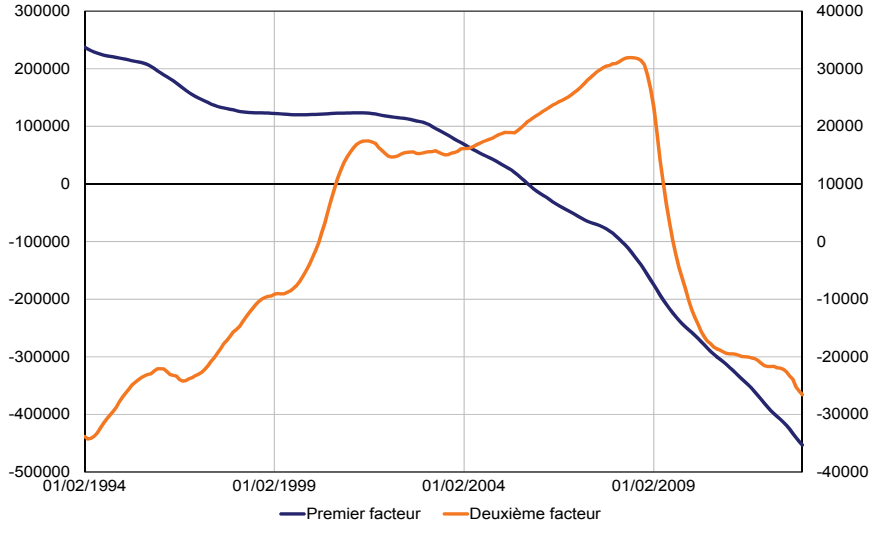
Graphique 1 : calcul des facteurs à partir de la matrice des moments non centrés d'ordre 2 sur variables non stationnaires seulement (méthode de Bai)



Source : DG Trésor.

- Variante : la méthode proposée par Banerjee (2009), consiste à cumuler les variables stationnaires avant de les adjoindre à la base des variables non stationnaires, pour appliquer la même approche à un ensemble de variables non stationnaires.

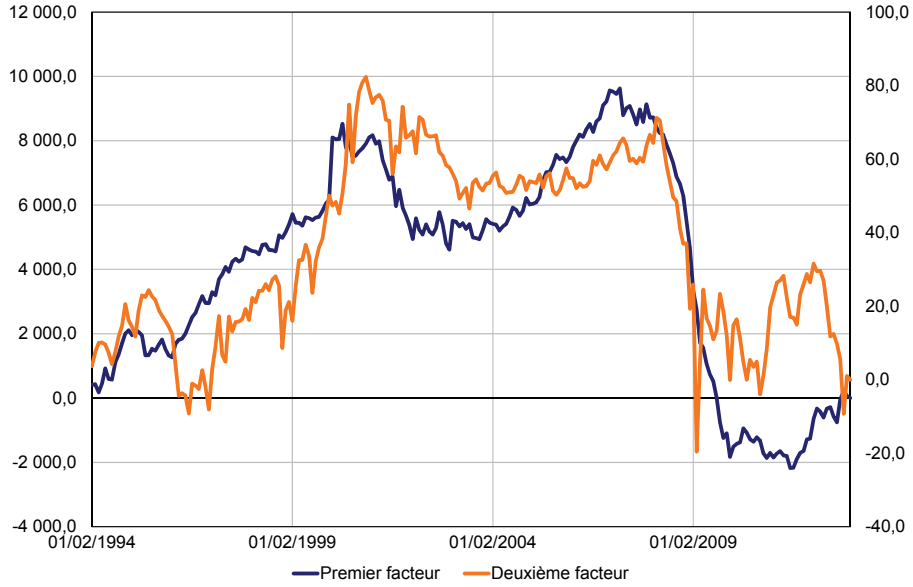
Graphique 2 : illustration de la méthode de Banerjee



Source : DG Trésor.

- Cas n°2 : Bai et Ng (2004) proposent au contraire de réaliser l'ACP sur des transformations stationnaires des variables non stationnaires et de cumuler ensuite les facteurs obtenus par ACP sur cette base. Cette approche permet, comme expliqué précédemment, de ne pas faire l'hypothèse de stationnarité des composantes idiosyncratiques.

Graphique 3 : illustration de la méthode de Bai et Ng (2004)



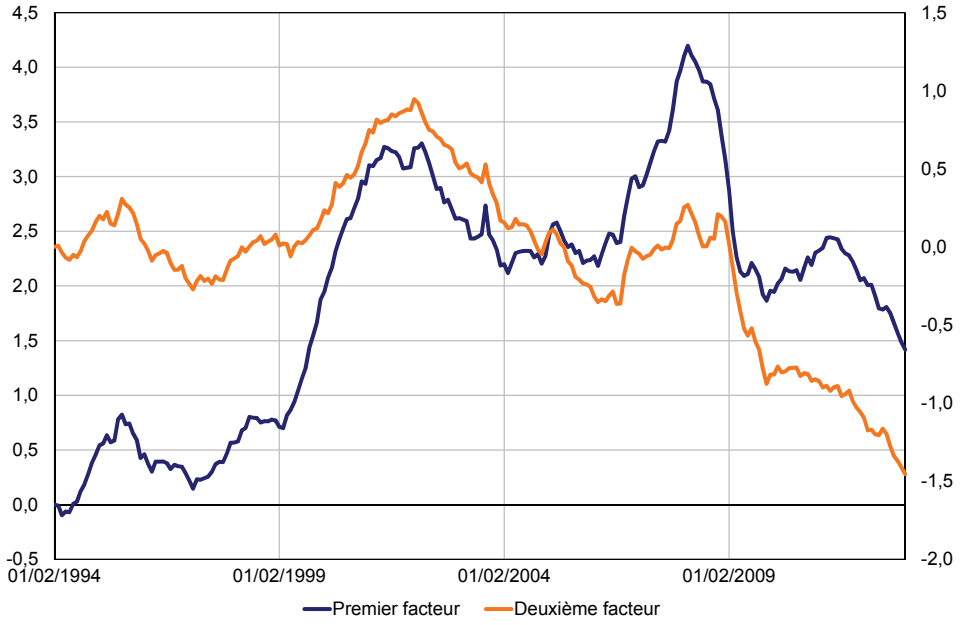
Source : DG Trésor.

La variante de la méthode utilisée dans le premier cas permet, en intégrant les variables stationnaires sous un format comparable aux variables non stationnaires (en les cumulant), d'obtenir un premier facteur assimilable à une tendance stochastique commune (il est bien intégré d'ordre 1), mais cette approche ne semble pas robuste sur le plan théorique : il y a peu de chances que les variables stationnaires cumulées pour l'occasion vérifient l'hypothèse de stationnarité des composantes idiosyncratiques.

Les facteurs issus du cas n°2 sont obtenus par cumul *ex post* des facteurs issus de l'ACP sur données stationnarisées. On voit mal comment ils pourraient correspondre aux seules tendances stochastiques communes de la théorie de la cointégration, et pourquoi leur nombre devrait correspondre au nombre des facteurs communs aux variables stationnarisées. De plus, en procédant ainsi on somme toutes les erreurs d'estimation.

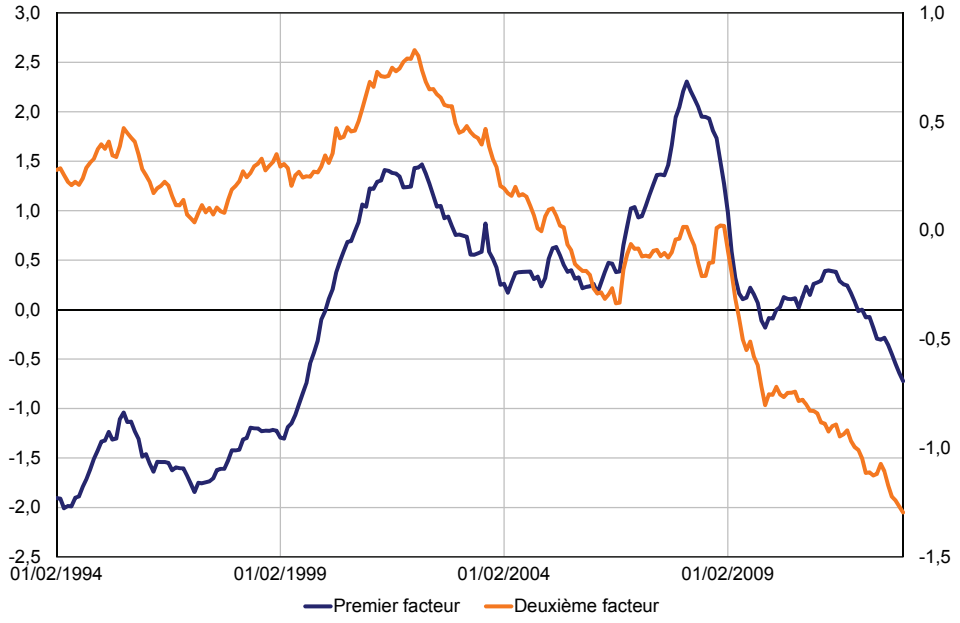
Une alternative à la méthode proposée par Bai qui pourrait permettre de compenser en partie les problèmes d'échelle, serait de prendre les variables en écart par rapport à leurs valeurs initiales, mais on constate que les facteurs obtenus sont proches de ceux obtenus par la méthode de Bai (ils ne sont cependant pas dans le même ordre), ou d'une simple ACP sur matrice de variance covariance empirique.

Graphique 4 : calcul des facteurs à partir de la matrice des moments d'ordre 2 sur variables stationnaires et non stationnaires prises en écart à leur valeur initiale



Source : DG Trésor.

Graphique 5 : ACP sur variables non stationnaires seulement



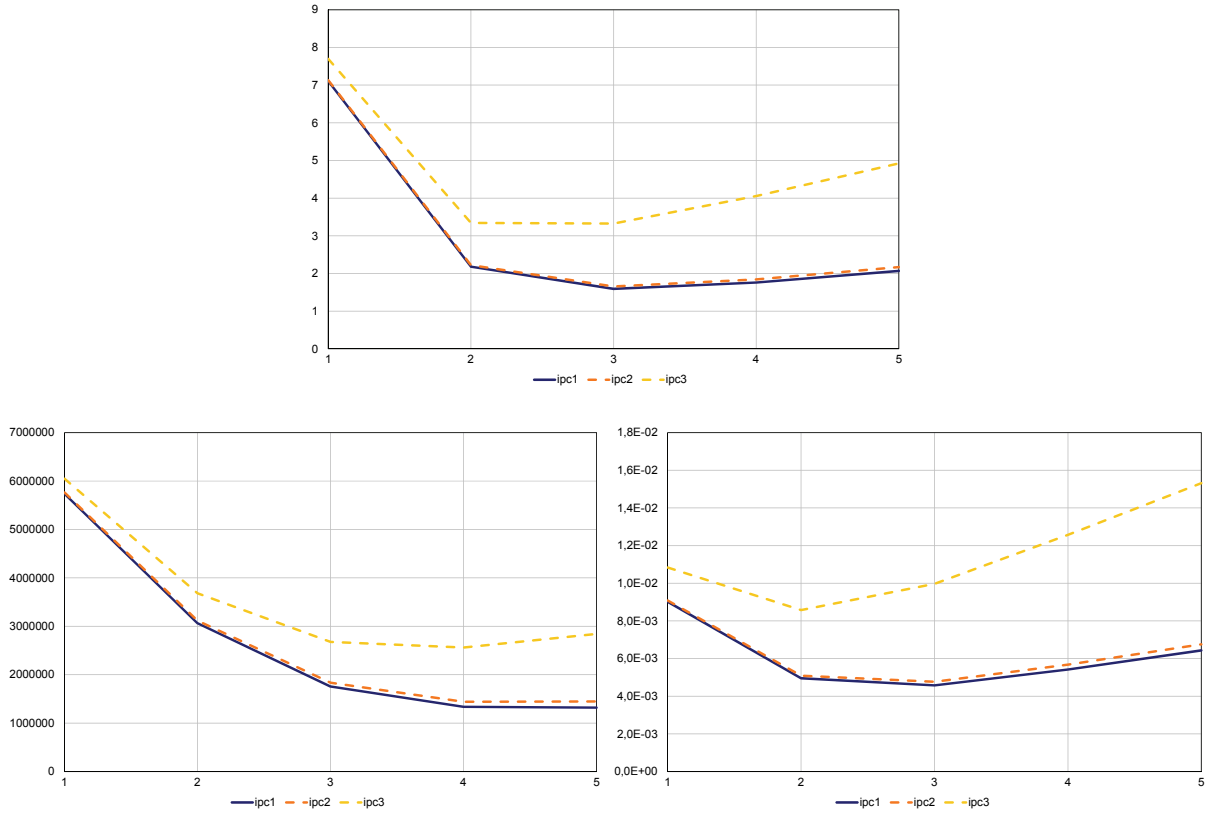
Source : DG Trésor.

Annexe 3 : Détermination du nombre de facteurs non stationnaires à conserver

Les critères d'information de Bai décrits en 1.2.4 semblent indiquer que 2 à 3 facteurs pourraient suffire.

Cependant, il faut noter que ces critères conduisent à des conclusions différentes en fonction de la méthode d'estimation des facteurs non stationnaires retenue, et qu'ils sont en outre sensibles à la valeur maximale kmax choisie *a priori* pour le nombre de ces facteurs.

Graphique 1 : critères IPC de Bai (2004) appliqués à l'approche de Bai (en haut), à l'approche de Banerjee (en bas à gauche), à l'ACP (en bas à droite)



Source : DG Trésor.

L'algorithme d'Onatski (2009) qui présente en théorie l'avantage de fournir des résultats plus précis, indique qu'il est possible de ne retenir qu'un facteur.

Cependant, d'après les tests de racine unitaire sur les résidus de la régression sur deux facteurs des variables I(1), on trouve des résidus stationnaires dans la très grande majorité des cas, ce qui n'est pas le cas si on ne conserve qu'un facteur.

Dans l'absolu, il faudrait refaire l'ensemble de ces tests à mesure que de nouvelles informations sont disponibles, on voit bien que ce n'est pas possible en pratique. Au vu de ces résultats sur l'échantillon complet, on décide donc de fixer à deux le nombre de facteurs non stationnaires.

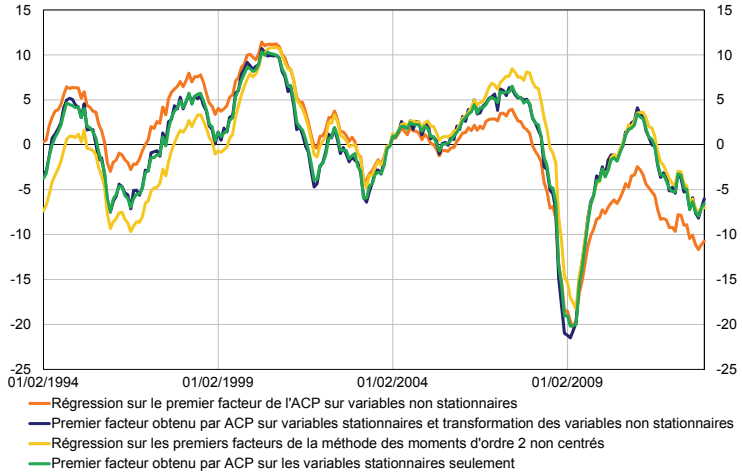
Annexe 4 : Traitement des variables stationnaires

Pour le traitement des données stationnaires, on peut envisager plusieurs approches :

- On se restreint aux variables stationnaires à titre de comparaison ;
- On adjoint à la base des variables stationnaires les transformations stationnaires des variables non stationnaires (ici on effectue une simple différenciation de ces variables) ;
- On régresse les variables I(1) sur les facteurs non stationnaires retenus en première étape (par la méthode des moments non centrés d'ordre 2), et on adjoint les résidus issus de cette régression, sous l'hypothèse que ceux-ci soient bien stationnaires.

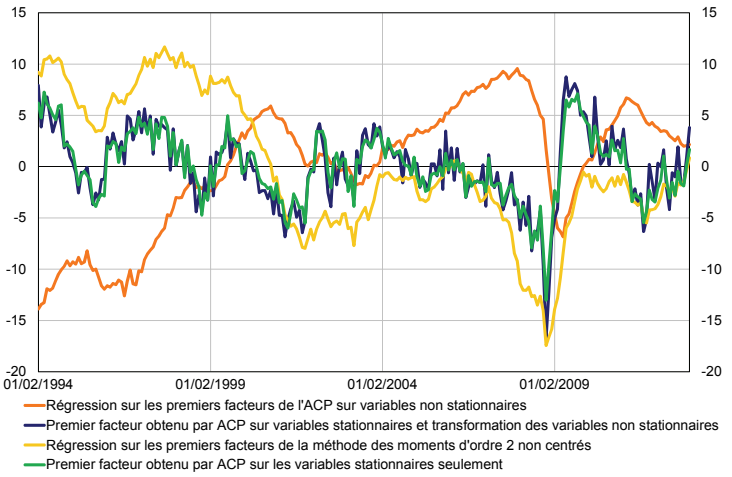
Nous n'avons pas retenu l'approche de Banerjee qui consistait à isoler les résidus pour calculer un facteur stationnaire commun à ces derniers.

Graphique 1 : comparaison du premier facteur de l'ACP sur variables stationnaires au premier facteur de l'ACP sur variables stationnaires et résidus de la régression des variables non stationnaires sur les premiers facteurs non stationnaires



Source : DG Trésor.

Graphique 2 : comparaison du deuxième facteur de l'ACP sur variables stationnaires au deuxième facteur de l'ACP sur variables stationnaires et résidus de la régression des variables non stationnaires sur les premiers facteurs non stationnaires



Source : DG Trésor.

Il est possible de calculer les contributions des variables aux facteurs et on peut ainsi voir que les variables I(1) transformées (que ce soit par différenciation ou sous forme de résidus) interviennent bien dans le calcul du premier facteur, ce qui justifie de les considérer en deuxième étape.

Annexe 5 : ACP et détermination des tendances communes

Les simulations sont réalisées suivant le protocole suivant :

$$x_{it} = \sum_{j=1}^{r} \Lambda_{ij} f_{jt} + e_{it}$$

$$\boldsymbol{x_t} = \Lambda \boldsymbol{f_t} + \boldsymbol{e_t}, \qquad \text{en notation vectorielle}$$

$$A(L)\boldsymbol{f_t} = \boldsymbol{u_t}, \qquad \text{avec } \boldsymbol{u_t} \text{ i. i. d. } N(0, I_r)$$

$$D(L)\boldsymbol{e_t} = \boldsymbol{v_t}, \qquad \text{avec } \boldsymbol{v_t} \text{ i. i. d. } N(0, \tau)$$

$$A_{ij}(L) = \begin{cases} 1 - L & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}, i, j = 1, \dots, r$$

$$D_{ij}(L) = \begin{cases} 1 - dL & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}, i, j = 1, \dots, r$$

$$\Lambda_{ij}\ i.i.d.N(0,1) \quad \text{pour} \quad i = 1, \dots, n, \qquad j = 1, \dots, r$$

$$\alpha_i = \frac{\beta_i}{1-\beta_i} \frac{T}{3} \sum_{j=1}^{r} \Lambda_{ij}^2 \text{ avec } \beta_i \textit{ i.i.d.} \text{ dans une loi uniforme } U[u; 0{,}5 - u],$$

Ce terme contrôle l'ordre de grandeur du ratio de la variance empirique de e_{it} par rapport à celle de x_{it} pour que sur l'échantillon étudiée la variance des premiers ne soient pas trop négligeables devant celle de ces derniers[13].

$$\tau_{ij} = \tau^{|i-j|}(1 - d^2)\sqrt{\alpha_i \alpha_j}$$

Ces termes contrôlent les corrélations des composantes idiosyncratiques

Il est également possible de simuler des facteurs stationnaires, dans ce cas on prend :

$$A_{ij}(L) = \begin{cases} 1 - d'L & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}, i, j = r + 1, \dots, R$$

Si on veut considérer l'impact de la présence de variables stationnaires dans la base, on prendra les $\Lambda_{ij} = 0$ pour $j = 1, \dots, r$ pour x_{it} variable stationnaire.

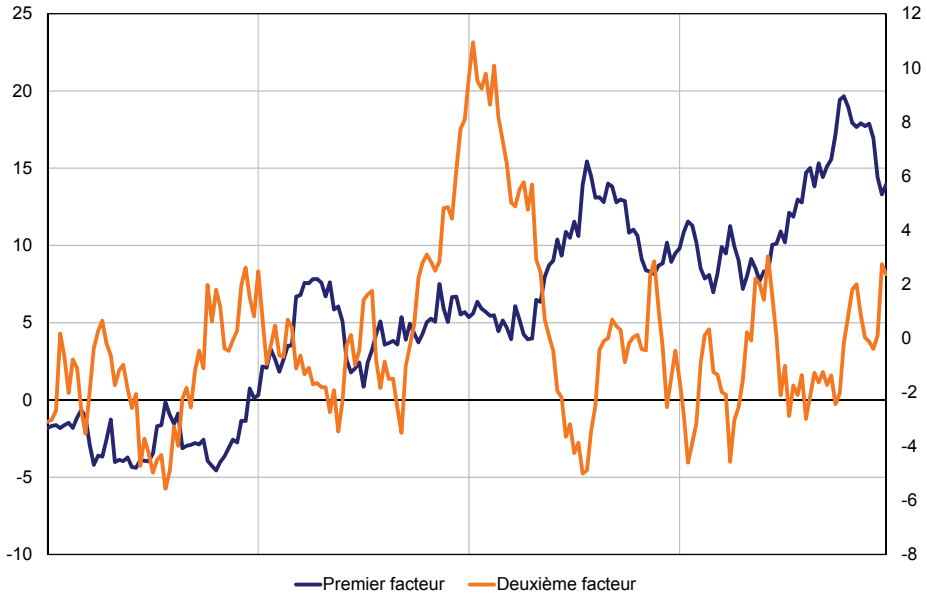
Pour prendre en considération la présence d'ordres de grandeur de variances de x_{it} différents, on peut également tirer$\Lambda_{ij}\ i.i.d.N(0, \sigma_i)$ où σ_i est elle-même tirée dans une loi uniforme. Mais c'est surtout en donnant plus de latitude au paramètre β_i, que les résultats sont fortement impactés.

$$N(0, {\sigma_i}^2)$$

Dans un premier temps, on considère un cas simple où les variables sont toutes I(1) et admettent deux facteurs communs I(1).

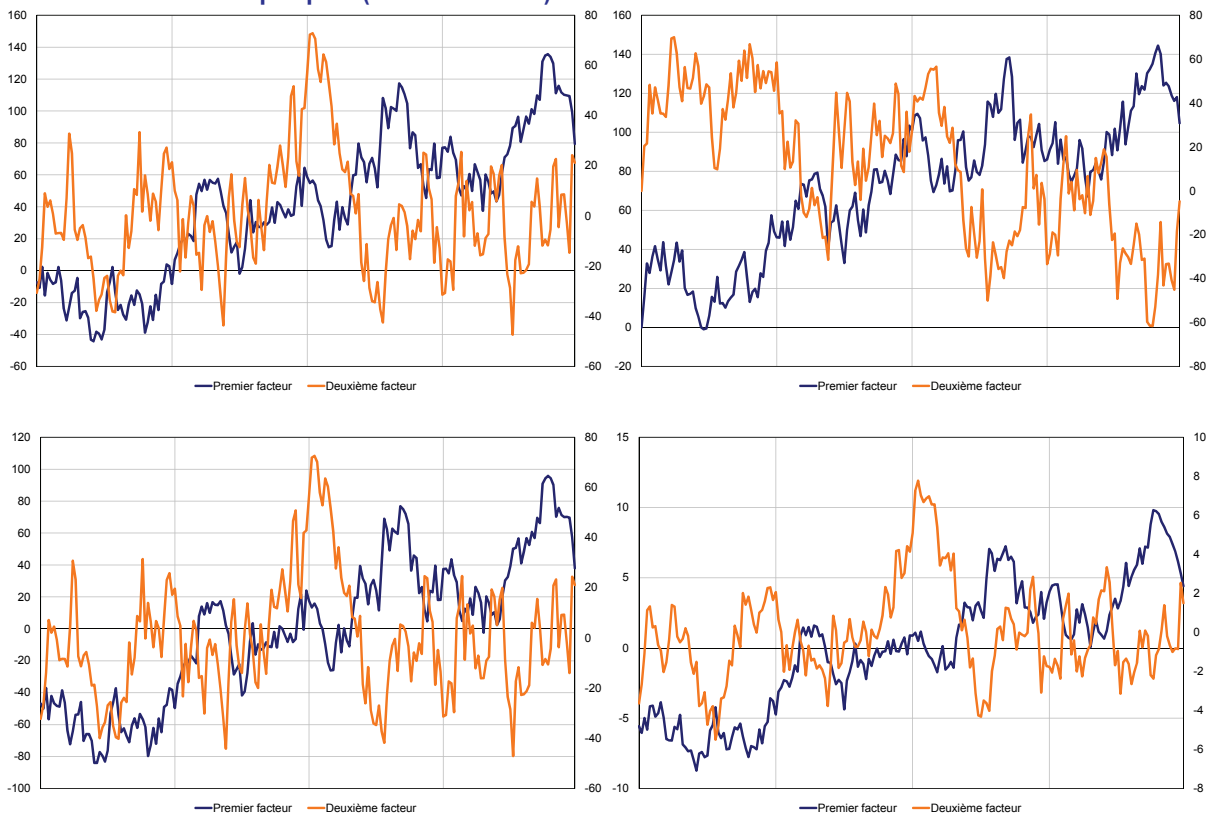
[13] Pour une interprétation de ces différents paramètres dans le cadre stationnaire voir Doz, Gianonne et Reichlin (2012).

Graphique 1 : facteurs simulés



Source : DG Trésor.

Graphiques 2 : facteurs estimés par la méthode des moments non centrés d'ordre 2 (en haut à gauche), par la méthode des moments d'ordre 2 appliquée aux variables prises en écart à leur valeur initiale (en haut à droite), par ACP sur matrice de variances-covariances empiriques (en bas à gauche), sur matrice d'autocorrélations empiriques (en bas à droite)

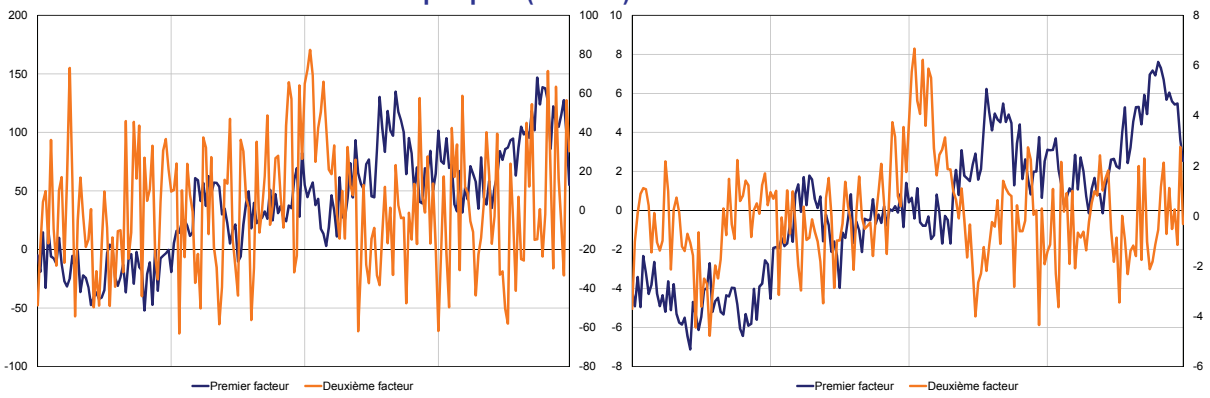


Source : DG Trésor.

Si on rajoute de l'autocorrélation, on obtient des estimations sensiblement similaires pour les différentes méthodes.

On peut aussi regarder ce qui se passe en jouant sur le paramètre β_i, afin de voir l'impact sur l'estimation de la présence de variables de variances très différentes, ce que nous avons effectivement observé dans la pratique. On voit que les facteurs estimés, d'autant plus sur données non centrées réduites, sont beaucoup plus volatils que les facteurs simulés. D'où l'importance de gérer les problèmes d'échelle.

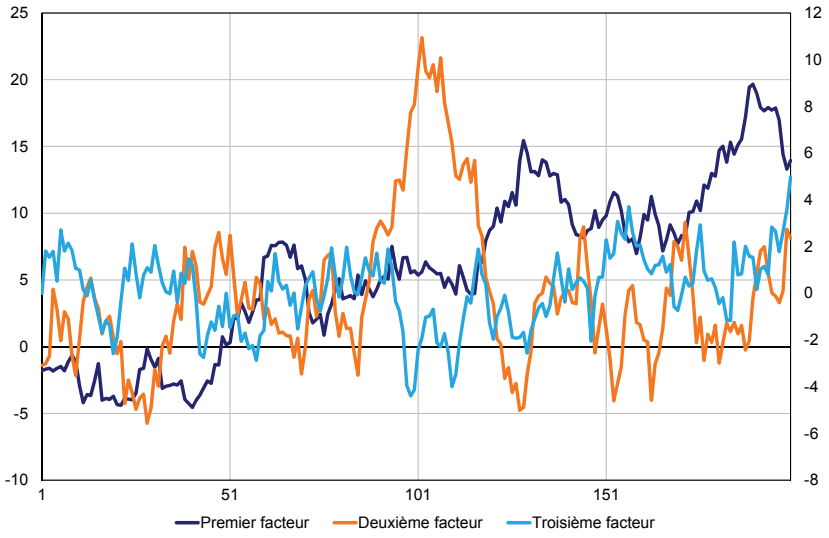
Graphiques 3 : facteurs estimés par la méthode des moments non centrés d'ordre 2 (à gauche), par ACP sur la matrice d'autocorrélations empiriques (à droite)



Source : DG Trésor.

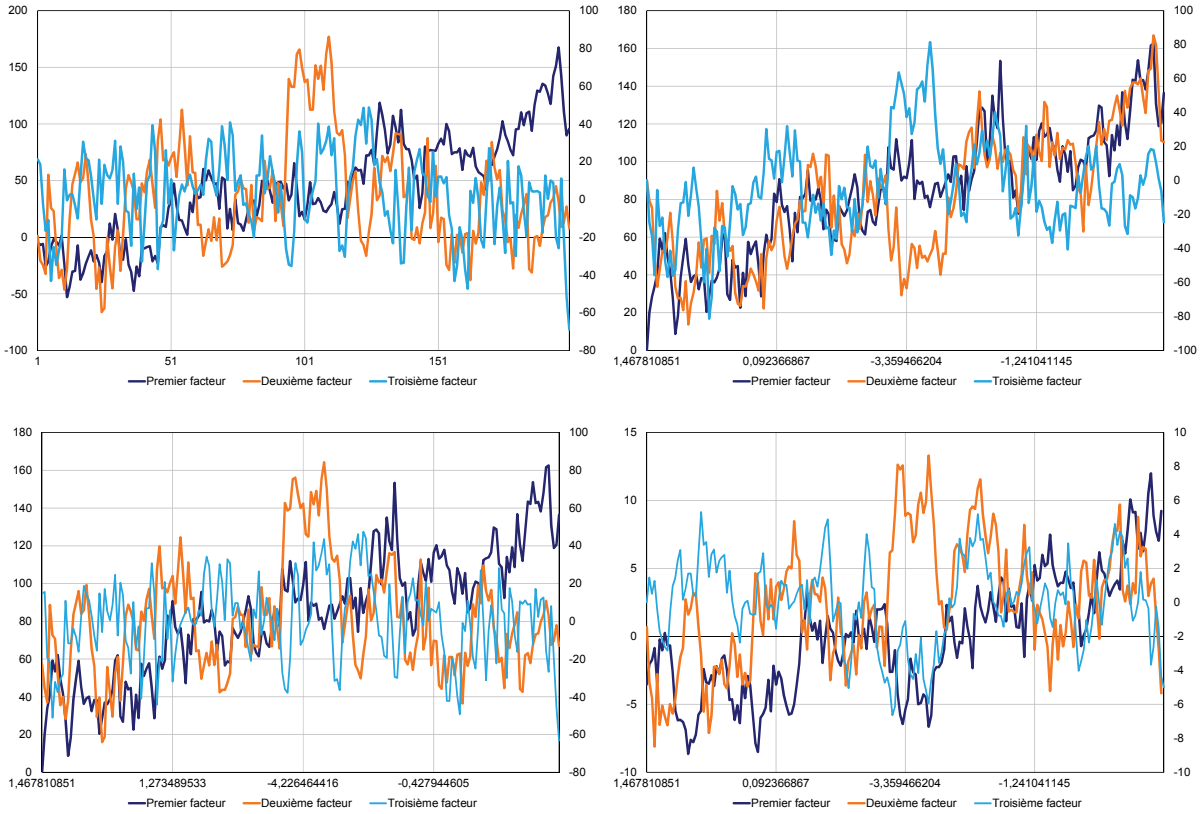
S'agissant de la question du traitement des variables stationnaires séparément ou conjointement aux variables non stationnaires, les simulations tendent à suggérer qu'il est plus pertinent de les séparer. En considérant un facteur stationnaire commun à l'ensemble des données en plus des deux facteurs non stationnaires et une centaine de variables stationnaires en plus des variables non stationnaires, nous obtenons les estimations suivantes :

Graphique 4 : facteurs simulés



Source : DG Trésor.

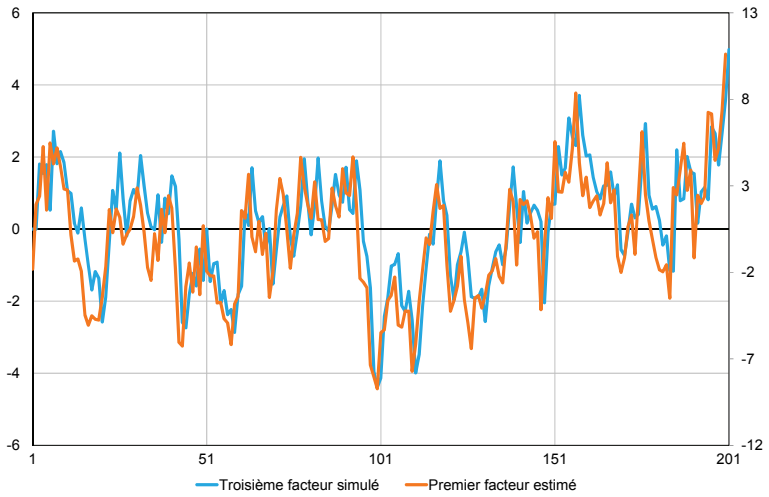
Graphiques 5 : facteurs estimés par la méthode des moments non centrés d'ordre 2 (en haut à gauche), par la méthode des moments d'ordre 2 appliquée aux variables prises en écart à leur valeur initiale (en haut à droite), par ACP sur matrice de variances-covariances empiriques (en bas à gauche), sur matrice d'autocorrélations empiriques (en bas à droite)



Source : DG Trésor.

On voit notamment que dans la méthode des moments d'ordre deux calculés sur les variables prises en écart à leur valeur initiale, le 3e facteur se rapproche du 2e facteur simulé et estimé dans les autres méthodes, « l'ordre » intuitif des facteurs (*cf.* 1.4) n'est donc pas respecté sur un nombre restreint d'observations. C'est un premier argument pour traiter séparément les variables non stationnaires et les variables stationnaires. De plus, le 3e facteur semble également non stationnaire. Les composantes stationnaires et non stationnaires ne sont donc pas clairement distinguées par cette approche. L'estimation du facteur stationnaire sur les seules données stationnaires semble beaucoup plus efficace.

Graphique 6 : facteurs estimés par ACP sur la matrice d'autocorrélation des seules variables stationnaires



Source : DG Trésor.

Annexe 6 : Équation de prévision du FECM d'après l'approche de Johansen sous l'hypothèse d'exogénéité faible

Si l'on souhaite obtenir une estimation plus rigoureuse et plus robuste que celle que l'on obtient à partir de l'approche en deux étapes, on estime un modèle vectoriel (ou FECM). Le nombre de facteurs non stationnaires et stationnaires sollicités dans cette démarche ainsi que leurs retards peuvent nous conduire à estimer un trop grand nombre de coefficients.

Sous hypothèse d'exogénéité faible des facteurs stationnaires, on peut cependant ne retenir que la première équation du vecteur à correction d'erreur estimé par la méthode de Johansen.

Supposons qu'on a un vecteur $Z_t = \begin{pmatrix} z_t \\ f_t^0 \end{pmatrix}$ avec $z_t \sim I(1)$ cointégré et $f_t^0 \sim I(0)$, avec ici $z_t = \begin{pmatrix} y_t \\ f_t^1 \end{pmatrix}$ et f_t^1 et f_t^0 les facteurs non stationnaires et stationnaires respectivement.

En toute rigueur, le modèle VECM associé peut s'écrire sous la forme suivante :

$$\begin{pmatrix} \Delta z_t \\ \Delta f_t^0 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} - \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} \beta' & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} z_{t-1} \\ f_{t-1}^0 \end{pmatrix} + \sum_{k=1}^{p-1} \begin{pmatrix} c_{11}^k & c_{12}^k \\ c_{21}^k & c_{22}^k \end{pmatrix} \begin{pmatrix} \Delta z_{t-k} \\ \Delta f_{t-k}^0 \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix}$$

Le terme à correction d'erreur s'écrit :

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} \beta' z_{t-1} \\ f_{t-1}^0 \end{pmatrix} = \begin{pmatrix} \alpha_{11}\beta' z_{t-1} + \alpha_{12} f_{t-1}^0 \\ \alpha_{21}\beta' z_{t-1} + \alpha_{22} f_{t-1}^0 \end{pmatrix}$$

On ne peut estimer β à partir de l'équation de Δz_t que si f_t^0 est faiblement exogène pour β, c'est-à-dire $\alpha_{21} = 0$. Si c'est le cas, alors on peut décorréler les deux équations en prémultipliant le système par $\begin{pmatrix} I & -A \\ 0 & I \end{pmatrix}$ où $A = \Sigma_{12}\Sigma_{22}^{-1}$ avec $V(\epsilon_t) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ et dans ce cas la première équation est remplacée par la première composante du vecteur :

$$\Delta z_t = A\,\Delta f_t^0 + \widetilde{\mu_1} + \widetilde{\alpha_{11}}\beta' z_{t-1} + \widetilde{\alpha_{12}} f_{t-1}^0 + \sum_{k=1}^{p-1} \widetilde{c_{11}^k}\Delta z_{t-k} + \sum_{k=1}^{p-1} \widetilde{c_{12}^k}\Delta f_{t-k}^0 + \widetilde{\epsilon_{1t}}$$

Cette équation peut être facilement réécrite de sorte à faire apparaître les f_{t-k} plutôt que les Δf_{t-k}.

Sous hypothèse d'exogénéité faible, les équations de prévision (4a) et (3a) utilisées pour les différentes méthodes décrites précédemment deviennent respectivement, en ne conservant que la première composante du vecteur Δz_t :

$$\Delta y_{t+h} = A_1\,\Delta f_t^0 + \tilde{\mu}_{1,1} + \tilde{\alpha}_{11,1}\beta' z_{t-1} + \sum_{k=1}^{p-1} \tilde{c}_{11,1}^k \Delta z_{t-k} + \sum_{k=1}^{p} \tilde{d}_{12,1}^k f_{t-k}^0 + \tilde{\epsilon}_{t+h} \quad \text{(3a bis)}$$

Et :

$$\Delta y_{t+h} = A_1\,\Delta f_{t+h}^0 + \tilde{\mu}_{1,1} + \tilde{\alpha}_{11,1}\beta' z_{t-1} + \sum_{k=1}^{p-1} \tilde{c}_{11,1}^k \Delta z_{t-k} + \sum_{k=1}^{p} \tilde{d}_{12,1}^k f_{t+h-k}^0 + \tilde{\epsilon}_{t+h} \quad \text{(4a bis)}$$

Annexe 7 : Modèle espace-état et filtre de Kalman

Cette annexe décrit la représentation espace état, puis le filtre de Kalman qui permet d'estimer les variables cachées (les facteurs dans notre application) à tout instant *t* conditionnellement aux variables observées jusqu'à *t*. L'algorithme de lissage permet ensuite d'affiner l'estimation en prenant en compte à tout instant l'ensemble des observations de Y[14].

1. Modèle espace-état, filtre et lisseur de Kalman

1.1 Le modèle espace - état

Le modèle espace-état du processus multivarié Y_t est représenté par le système d'équations (1) et (2) suivant :

$$\begin{cases} \alpha_t = A\alpha_{t-1} + c + \varepsilon_t \quad (1) \\ Y_t = C\alpha_t + e_t \quad (2) \end{cases} \qquad \begin{pmatrix} \varepsilon_t \\ e_t \end{pmatrix} \approx BB\left(0, \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}\right)$$

avec $EZ_0 = \mu, VZ_0 = P$. L'équation (1) est l'équation d'état et l'équation (2) est l'équation de mesure.

On appelle par ailleurs :

- Y_t la variable de mesure à la date *t*
- α_t la variable d'état à la date *t*
- ϵ_t le vecteur d'innovations à la date *t*
- e_t le vecteur des erreurs de mesure à la date *t*
- A la matrice de transition
- C la matrice de mesure
- $C\alpha_t$ le signal à la date *t*

On considère ici le système sous sa forme canonique car on suppose que :

$$E(\varepsilon_t e_s) = E(\varepsilon_t \alpha_0) = E(e_t \alpha_0) = 0 \qquad \forall t, s = 1, \ldots, T$$

On suppose ici par ailleurs que les matrices A et C ne dépendent pas du temps (le système est dit invariant par rapport au temps).

1.2 Le filtre de Kalman

Le filtre de Kalman permet d'estimer à chaque instant *t* le vecteur d'état conditionnellement à la variable *Y* observée jusqu'à la date *t*, soit $\alpha^*_{t|t} = E(\alpha_t | Y_0, \ldots, Y_t)$, *t=1,…,T*. Cet algorithme consiste à itérer les cinq étapes suivantes pour *t=1*,…,*T* :

[14] On trouvera plus de détails et des éléments de démonstration par exemple dans le chapitre 15 de l'ouvrage de C. Gourieroux et A. Monfort, « Séries temporelles et modèles dynamiques ».

$$\begin{cases} (1)\ \alpha^*_{t|t} = \alpha^*_{t|t-1} + K_t\left(Y_t - C\alpha^*_{t|t-1}\right) \\ \text{où}\ K_t = P_{t|t-1}C'(CP_{t|t-1}C'+R)^{-1} \\ (2)\ P_{t|t} = (I - K_tC)P_{t|t-1} \\ (3)\ \alpha^*_{t+1|t} = A\alpha^*_{t|t} + c \\ (4)\ P_{t+1|t} = AP_{t|t}A'+Q \end{cases}$$

avec $\alpha^*_{t|t} = E(\alpha_t|Y_0, \dots, Y_t)$, l'estimation courante du vecteur d'état, $\alpha^*_{t+1|t} = E(\alpha_{t+1}|Y_0, \dots, Y_t)$, la prévision de α_t pour l'instant t+1 faite en t et $\mathrm{P}_{t+1|t} = V(\alpha_t - \alpha_{t+1|t})$, $\mathrm{P}_{t|t} = V\left(\alpha_t - \alpha_{t|t}\right)$ la variance de l'erreur de prévision. La matrice K_t, est appelée matrice de gain. Les deux premières équations actualisent l'estimation de α_t, (équation 1) et la précision correspondante (équation 2). Dans l'équation (1), l'estimation courante du vecteur d'état est obtenue à partir de l'estimation faite à la date *t-1*, à laquelle on ajoute un terme prenant en compte l'information nouvelle contenue dans l'observation de Y_t ; ce terme, qui se déduit d'un programme de minimisation de la variance de l'erreur courante, s'exprime simplement en fonction de l'estimation à la date *t-1* de α_t, de la variable de mesure Y_t, et de la matrice de gain K_t, définie par l'équation (5). Les équations (3) et (4) mettent à jour l'estimation de la variable d'état en *t+1* à partir de l'équation d'état.

L'initialisation de l'algorithme nécessite la connaissance de μ et de P. En effet, $\alpha^*_{1|0} = \mu$, et $P_{1|0} = P$. Il faut donc avoir un a priori sur α_0 pour que le processus converge. Lorsque le processus α_t est stationnaire (comme c'est le cas ici) *μ* et *P* peuvent être explicitement calculés à partir des paramètres du modèle.

1.3 L'algorithme de lissage

L'algorithme de lissage permet ensuite d'affiner l'estimation des états en conditionnant l'estimation non pas à la seule information présente et passée mais à l'ensemble de l'information. On itère cette fois-ci en arrière les calculs suivants pour *t=T-1* à 1.

$$\begin{cases} \alpha^*_{t|T} = \alpha^*_{t|t} + F_t\left(\alpha^*_{t+1|T} - \alpha^*_{t+1|t}\right) \ \text{ avec } F_t = P_{t|t}A'P^{-1}_{t+1|t} \\ P_{t|T} = P_{t|t} + F_t\left(P_{t+1|T} - P_{t+1|t}\right)F_t' \end{cases}$$

Les calculs sont initialisés en *T* par les produits du filtre.

L'ensemble des calculs précédents (filtre et lissage) suppose que les matrices A, C, Q et R sont connues. Si tel n'est pas le cas, la log-vraisemblance du modèle peut être construite au cours des itérations du filtre et l'on pourra estimer les paramètres inconnus par maximum de vraisemblance. Dans notre application, on utilise comme matrices A et C les estimateurs de la première étape et comme matrices Q et R les matrices de variance covariance des résidus associés.

2. Cadre non stationnaire

On suppose $F_t \sim I(1)$ non cointégré et on considère qu'il suit un VAR d'ordre 2, on peut écrire : $\Phi(L)F_t = \mu + \epsilon_t$

on peut donc poser $\Phi(L) = \Phi(1) + (1-L)\Phi^*(L)$.

$$\Phi(1)F_t + (1-L)\Phi^*(L)F_t = \epsilon_t$$

Si F_t est non cointégré on a $\Phi(1) = 0$, $\Phi(L) = (1-L)\Phi^*(L)$, et on peut écrire :

$$(I-L)(I-\Phi L)F_t = \mu + \epsilon_t$$

On pose : $(1-L)F_t = G_t$ où G_t est donc la différence première de F_t, et on suppose qu'elle suit un processus autorégressif d'ordre 1 : $(I-\Phi L)G_t = \epsilon_t$.

Le modèle espace-état s'écrit donc $\begin{cases} x_t = \Lambda F_t + e_t \\ F_t = F_{t-1} + G_t \\ G_t = \mu + \Phi G_{t-1} + \epsilon_t \end{cases}$

En posant $\alpha_t = \begin{pmatrix} F_t \\ G_{t+1} \end{pmatrix}$, on retombe sur les expressions présentées en 1.1, avec l'équation de transition :

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

et l'équation de mesure : $x_t = (\Lambda \quad 0)\begin{pmatrix} F_t \\ G_{t+1} \end{pmatrix} + e_t$

Se pose la question de l'initialisation, avec $\alpha_1 = \begin{pmatrix} F_1 \\ G_2 \end{pmatrix}$, à la différence qu'ici VF_1 n'est pas fini.

On note $P_1 = \begin{pmatrix} VF_1 & cov(F_1, G_2) \\ cov(F_1, G_2) & VG_2 \end{pmatrix}$

On calcule VG_2 classiquement en utilisant $VG_t = \Gamma_0$, avec $(1 - \Phi\otimes\Phi)vec\Gamma_0 = vec\Sigma$ où $\Sigma = V\epsilon_t$

On initialise VF_1 par $+\infty$. Pour cela on prend $VF_1 = \kappa I$ avec κ grand, en supposant que que les composantes de F_1 sont non corrélées.

Par ailleurs, on peut écrire : $F_1 = F_{-H} + \sum_{s=-H+1}^{1} G_s$.

On en déduit : $Cov(F_1, G_2) = Cov(F_{-H}, G_2) + \sum_{s=-H+1}^{1} Cov(G_s, G_2)$

$$= Cov(F_{-H}, G_2) + \sum_{s=-H+1}^{1} \Gamma_G(s-2)$$

$$= Cov(F_{-H}, G_2) + \sum_{s=-H+1}^{1} \Gamma'_G(2-s)$$

$$= Cov(F_{-H}, G_2) + \sum_{h=1}^{H+1} \Gamma'_G(h)$$

Comme $G_t = \mu + \Phi G_{t-1} + \epsilon_t$, $\Gamma_G(h) = \Phi\Gamma_G(h-1) + 0, \forall\, h > 0$

$$\Rightarrow \Gamma_G(h) = \Phi^h \Gamma_G(0)$$

Par suite :

$$Cov(F_1, G_2) = Cov(F_{-H}, G_2) + \sum_{h=1}^{H+1} \Gamma_G(0)\Phi'^h$$

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

$$= Cov(F_{-H}, G_2) + \Gamma_G(0)\left(\Phi\sum_{h=0}^{H} \Phi^h\right)'$$

On a $(I-\Phi)\sum_{h=0}^{H}\Phi^h = I-\Phi^{H+1} \xrightarrow[H\longrightarrow+\infty]{} I$. Donc $\sum_{h=0}^{H}\Phi^h \xrightarrow[H\longrightarrow+\infty]{} (I-\Phi)^{-1}$

$Cov(F_1 G_2') \simeq 0 + \Gamma_G(0)(I-\Phi')^{-1}\Phi'$quand $H \longrightarrow +\infty$

On aurait donc :

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

Références :

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

Bai J., Ng S. (2002), “Determining the Number of Factors in Approximate Factor Models”, *Econometrica,* Vol. 70, n°1, 191-221.

Bai J., Ng S. (2004), “A PANIC Attack on Unit Roots and Cointegration”, *Econometrica,* Vol. 72, n°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.

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

Banerjee A., Marcellino M. (2008), “Factor-augmented Error Correction Models”, CEPR *Discussion Papers* 6707.

Banerjee A., Marcellino M., Masten I. (2010), “Forecasting with Factor-augmented Error Correction Models”, *CEPR Discussion Papers* 7677.

Barhoumi K., Darné O. et Ferrara L. (2012), « Une revue de la littérature des modèles à facteurs dynamiques », *Économie et Prévision,* n°199.

Barigozzi M., Lippi M., Luciani M. (2013), “Generalized Dynamic factor Models, Cointegration and Error Correction Mechanisms”, *Mimeo*.

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. et Doz C. (2011), « Prévision de court terme de la croissance du PIB français à l’aide de modèles à facteurs dynamiques », *Documents de travail de la DG Trésor,* Numéro 2011/01.

Bessec, M. et Doz C. (2012), « Prévision à court terme de la croissance du PIB français à l’aide de modèles à facteurs dynamiques », *Économie et Prévision,* n°199, 1-30.

Beveridge S., Nelson C. (1981), “A New Approach to Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to Measurement of the "Business Cycle"”, *Journal of Monetary Economics,* n°7, 151-174.

Charpin F. (2011), *«* Réévaluation des modèles d’estimation précoce de la croissance », *Revue de l’OFCE* n°108.

Combes S., Doz C. et Fournier J.M. (2013), « Prévision de court terme de la croissance du PIB français à l’aide de modèles à facteurs dynamiques : impact de la sélection de variables », *Documents de travail de la DG Trésor,* Numéro 2013/02.

Doz C., Giannone D. et Reichlin L. (2011), “A two-step estimator for large approximate dynamic factor models based on Kalman filtering”, *Journal of Econometrics* Vol. *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”, Estmation and Testing, *Econometrica*, Vol. 55, pp 251-276.

Engle R.F., Kozicki S. (1993), “Testing for common features”, *Journal of Business and Economic Statistics,* Vol. 11*, n°4.*

Escribano A., Pena D. (1994), “Cointegration and common factors”, *Journal of Time Series Analysis,* Vol. 15, n°6, 577-586.

Gourieroux C., Monfort A. (1990), « Séries temporelles et modèles dynamiques », *Economica*

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,* Vol. 92(4), *1004-1016.*

Stock J.H., Watson M. (1988), "Testing for Common Trends", *Journal of the American Statistical Association,* Vol. 83, n°404

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