Estimation précoce de la croissance

De la régression LARS au modèle à facteurs

Françoise Charpin
Université Paris II et
OFCE, Centre de recherche
en économie de Sciences Po

Dans ce travail, l'estimation précoce de la croissance provient d'un modèle à facteurs, extraits d'un nombre réduit de séries mensuelles, ces dernières ayant été choisies par l'algorithme de la régression LARS (Least Angle Regression). On suit en cela le travail de Bai et Ng (2008) qui tranche avec le traditionnel modèle à facteurs, basé sur un très grand nombre de séries mensuelles. Les auteurs préconisent de ne retenir que les séries les plus performantes pour prévoir la croissance « the targeted predictors ». Une pseudo analyse en temps réel est mise en œuvre sur la période 2001-2007 pour estimer la croissance française du trimestre en cours et du trimestre suivant.

francoise.charpin@ofce.sciences-po.fr

Mots clés : Prévision. Modèle à facteurs. Régression LARS.

Durant ces dernières années, les méthodes de prévision à court terme de la croissance se sont multipliées. La plupart repose sur un modèle à facteurs, ce qui désigne soit un modèle de régression dont les régresseurs sont des composantes principales, soit un modèle espace-état à composantes inobservables. Dans tous les cas, l'extraction des facteurs est basée sur un grand nombre de séries mensuelles publiées avant le PIB, séries qui coïncident avec l'activité ou qui sont avancées par rapport à elle.

Le premier type de modèle est issu des travaux de Stock et Watson (2002) et est bien connu maintenant. L'information collectée est résumée par les premières composantes principales de la matrice de corrélations des séries utilisées. Le modèle de prévision donne le taux de croissance du PIB réel en fonction des premières composantes principales courantes et retardées, la sélection de celles-ci se faisant selon les critères économétriques usuels. Cette approche a évolué à la suite des travaux de Bai et Ng (2002, 2006, 2008), Bai (2003) et de Boivin et Ng (2006). C'est précisément l'article de Bai et Ng (2008) qui est la base de notre travail. Le message principal des auteurs est qu'on obtient de meilleures prévisions en calculant les facteurs non pas à partir d'un grand nombre de séries, mais avec un nombre nettement plus réduit de séries bien choisies, ce choix reposant sur l'algorithme de la régression LARS (*Least Angle Regression*) ou de deux autres méthodes qui en sont des cas particuliers, la méthode LASSO (*Least Absolute Shrinkage and Selection Operator*) et la méthode EN (*Elastic Net*).

Le second type de modèle trouve son origine dans les travaux de Stock et Watson (1989), mais l'application à l'estimation précoce de la croissance résulte des travaux de Doz, Giannone et Reichlin (2006), avec des exemples dans Giannone, Reichlin et Small (2006). Cette dernière approche s'apparente en réalité à la précédente. Le modèle espace-état contient les équations d'état reliant toutes les séries mensuelles aux composantes inobservables, et des équations de transition, exprimant que les composantes inobservables suivent un processus vectoriel autorégressif. Pour estimer le modèle, les composantes inobservables sont initialisées par les premières composantes principales (dont il faut préciser le nombre *a priori*), puis ces dernières sont ensuite modifiées pour tenir compte de leur dynamique autorégressive (également donnée *a priori*). Le modèle est ainsi très proche d'un modèle de régression dont les régresseurs sont des composantes principales, auquel est adjoint un modèle VAR pour prévoir ces composantes et permettre ainsi une prévision à court terme de la croissance. Nous ne suivrons pas la voie des modèles à composantes inobservables, peu fiables pour obtenir des prévisions mensuelles car particulièrement instables quant à l'estimation de leurs paramètres, lorsque ces derniers sont nombreux. Cette voie a connu récemment un véritable engouement comme on a pu le constater durant le colloque ISF (*International Symposium on Forecasting*) 2008 où de nombreux papiers européens mettaient en œuvre cette méthode, à la suite des papiers de la BCE mentionnés ci-dessus.

Pour illustrer l'approche qui va être décrite, nous choisissons de travailler sur des données françaises. Nous allons calculer des estimations précoces de la croissance allant de moins 165 jours à moins 5 jours avant sa première publication. L'expérimentation est menée sur une période de 7 ans (2001-2007) et correspond à ce que l'on appelle usuellement une pseudo analyse en temps réel, c'est-à-dire que les données utilisées sont celles connues aujourd'hui et non celles auxquelles on aurait été confrontées dans le passé. De véritables données en temps réel ne sont pas disponibles pour la France, en revanche Eurostat développe actuellement des fichiers contenant ce type de données pour la zone

euro. La validation complète de la procédure proposée n'est donc pas assurée, car les données connaissent des révisions substantielles. Or, prévoir correctement la première estimation de la croissance ne relève peut-être pas de la même règle que prévoir correctement la croissance définitive ou semi-définitive. Prenons deux exemples. La production industrielle pourrait jouer un rôle plus important dans la première estimation de la croissance que celui qu'elle a réellement dans l'estimation du PIB finalement publiée. Les données d'enquêtes qui interviennent largement dans notre analyse pourraient avoir un rôle beaucoup plus faible en réalité si, par exemple, la première estimation de la croissance ne reposait pas, ou peu, sur ces données.

1. Quelle information considérer ?

À l'origine du modèle à facteurs se trouve l'idée que les fluctuations des agrégats macroéconomiques sont engendrées par un petit nombre de chocs (les facteurs) que l'on cherche alors à identifier pour mieux appréhender les évolutions conjoncturelles. Pour cela, on réunit le plus grand nombre possible de séries dont on extrait, par une méthode statistique, les fluctuations communes. Ces chocs peuvent être utilisés dans la construction d'indicateurs coïncidents, comme par exemple l'EuroCoin, ou peuvent permettre d'obtenir une estimation précoce de la croissance (pour le trimestre en cours), voire même une prévision à court terme (pour le trimestre suivant), comme dans Doz, Giannone et Reichlin (2006).

Mais d'un strict point de vue de la prévision d'une variable, utiliser le plus grand nombre possible de séries pour extraire des facteurs, même celles peu reliées à la variable d'intérêt, peut s'avérer nuisible parce que ces séries introduisent du bruit qui accroît la volatilité de la prévision. Ce point a été l'objet de l'article de Boivin et Ng (2006), qui mettent en évidence qu'utiliser des séries contenant peu d'information sur la variable à prévoir n'améliore pas la prévision. D'où l'idée de Bai et Ng (2008) de sélectionner un nombre restreint de séries, qui sont pertinentes pour l'objectif poursuivi, avant d'en extraire des facteurs de prévision. Ils s'intéressent donc dans leur article aux méthodes de sélection d'une information adaptée aux circonstances et conseillent l'utilisation de l'algorithme de la régression LARS (ou de ces deux cas particuliers, LASSO et EN). Il s'agit de sélectionner des séries qui apportent de l'information sur la variable à prévoir, séries qualifiées de « targeted predictors », alors qu'habituellement l'extraction des facteurs se faisait sans tenir compte de celle-ci. Ces séries « ciblées » peuvent être trop nombreuses pour être utilisées directement comme régresseurs dans une équation de prévision. En effet, la colinéarité des régresseurs empêche une estimation précise de leur impact respectif. D'où le passage par des facteurs, qui ici sont à choisir parmi les composantes principales de la matrice de corrélation des séries ciblées. Remarquons que l'utilisation des composantes principales était autrefois une manière de régler le problème de colinéarité (*Ridge Regression*). Le lien avec le passé peut aussi être fait à propos de la régression LARS, qui s'apparente à la régression pas à pas (*Stepwise Regression*), tout en étant plus générale et plus performante quant à notre objectif. L'esprit de la proposition de Bai et Ng, à savoir tenir compte de la grandeur d'intérêt avant d'extraire des facteurs, n'est pas non plus nouveau puisque c'est celui de la régression PLS (*Partial Least Squares*), et plus récemment, de la méthode décrite dans Bair, Hastie, Paul et Tibshirani (2006) (*Supervised Principal Component*).

2. Constituer une information ciblée

Pour cela, il faut une procédure de sélection de séries qui tienne compte de la variable à prévoir. Depuis les méthodes classiques de sélection (séquentielles ascendantes et descendantes), diverses méthodes sont apparues et, parmi elles, la sélection via the LASSO (Tibshirani, 1996), via the LARS (Efron, Hastie, Johnstone et Tibshirani, 2004), via the Elastic Net (Zou et Hastie, 2005).

Dans la sélection séquentielle ascendante (*forward selection regression*), la $(k+1)^e$ série entrante est celle qui présente une corrélation maximale (en valeur absolue) avec le résidu de la régression sur les k séries de la sélection précédente et l'ampleur du pas est donnée par cette corrélation. Cette méthode de sélection est jugée trop agressive car elle élimine trop de séries lorsque ces dernières sont corrélées avec celles incluses. Pour pallier ce défaut, il a été proposé de modifier l'ampleur du pas dans le cheminement précédent (*forward stagewise regression*). Mais cette procédure modifiée apparaît comme un cas particulier de la régression LARS sans présenter d'avantages particuliers.

Dans la sélection via the LARS, la $(k+1)^e$ série entrante, ainsi que l'ampleur du pas, sont déterminées de telle sorte que l'écart entre la série et son estimation à l'étape $(k+1)$ ait une corrélation identique avec les $(k+1)$ séries. Géométriquement, une corrélation est un angle et le cheminement (équi-angulaire) se fait selon un angle moindre (*least angle*) que celui qui aurait résulté de la corrélation de la série entrante avec le résidu de la régression sur les k séries.

La méthode LASSO correspond à une régression contrainte où la somme des carrés des erreurs est minimisée sous la contrainte que la somme des valeurs absolues des coefficients de régression reste inférieure à une borne C_1 donnée. Le calcul est effectué sur des séries préalablement centrées réduites (les coefficients de régression sont ainsi homogènes à des coefficients de corrélation). Plus C_1 est faible, moins il y a de variables dans la sélection. Les séries sélectionnées peuvent être ordonnées en examinant la valeur absolue du coefficient de régression. On peut montrer que cette méthode s'inscrit dans le cadre de la régression LARS, mais pour cela il faut en adopter une autre présentation (*cf.* Efron, Hastie, Johnstone et Tibshirani, 2004). On voit ainsi qu'en modifiant légèrement l'algorithme LARS, on obtient un algorithme LASSO désigné dans Bai et Ng (2008) par LARS/LASSO.

La présentation (en termes de minimisation sous contrainte) que nous avons évoquée, permet de voir que la régression LASSO s'apparente à la régression Ridge. En effet, dans la régression Ridge la contrainte porte sur la somme des carrés des coefficients de régression (plutôt que la somme de leurs valeurs absolues), somme qui doit rester inférieure à une borne C_2 donnée.

Ceci permet alors de présenter la méthode EN (*Elastic Net*), qui donne une solution intermédiaire entre les méthodes LASSO et Ridge. En effet, la sélection EN se fait en minimisant la somme des carrés des erreurs en plaçant deux contraintes, celles des méthodes LASSO et Ridge, avec des bornes C_1 et C_2 liées par une relation, telle que les deux cas extrêmes soient le cas LASSO et le cas Ridge. La méthode (EN) peut s'obtenir également en modifiant l'algorithme LARS (Bai et Ng parlent de LARS/EN) et est préconisée quand le nombre de séries initialement considérées est plus grand que le nombre d'observations (car alors, l'algorithme LARS/LASSO ne fonctionne pas).

Dans leur article, Bai et Ng (2008) utilisent un algorithme LARS/LASSO et, quand le nombre de séries dépasse le nombre d'observations, un algorithme LARS/EN. L'objectif y est la prévision du taux d'inflation à divers horizons et la comparaison de plusieurs méthodes, dont celles passant par des facteurs, soit construits en utilisant toutes les séries

disponibles (132), soit en utilisant un nombre plus restreint de séries ciblées (30). Deux autres types de modèles, qualifiés de non linéaires, sont aussi présentés. Dans un cas, les facteurs sont non-linéaires car construits en utilisant les séries et leur carré (264 séries), dans l'autre cas, les facteurs linéaires extraits des 132 séries peuvent entrer au carré dans l'équation de prévision, qui devient alors non linéaire. Ces modèles factoriels sont également comparés à deux modèles classiques de régression où un faible nombre de séries est directement utilisé pour prévoir l'inflation. Dans le premier modèle, on utilise systématiquement les 5 premières séries de la sélection LARS/LASSO ; dans l'autre, on utilise les premières séries de la sélection dont le nombre est arrêté à l'aide du critère BIC. L'article de Bai et Ng se termine en concluant que le modèle factoriel sur la base de séries ciblées est le plus prometteur.

3. Notre sélection pour la croissance française

Dans l'article de Bai et Ng (2008), la base de données est constituée de 132 séries [1], sans tenir compte du fait que certaines séries sont avancées, plutôt que coïncidentes, par rapport à la grandeur à prévoir. Bien sûr, cela pourrait être pris en compte par la suite, en retardant certains facteurs, mais la sélection préliminaire aura été faite sans tenir compte d'éventuelles avances, ce qui n'apparaît pas optimal. Au contraire, il paraît plus judicieux d'introduire dans la base de données la valeur courante et les valeurs retardées des séries (au moins pour certaines) et d'attendre de la méthode de sélection qu'elle renseigne sur le caractère avancé ou non d'une série. Autre remarque du même ordre : il arrive fréquemment avec les variables d'enquête que l'opinion ou/et la variation de l'opinion soient significatives dans une régression, c'est-à-dire le niveau ou/et la différence première d'une série [2]. Ainsi, en introduisant dans la base de données aussi bien le niveau que la différence première d'une variable, on pourra attendre de la méthode de sélection qu'elle retienne l'une ou l'autre forme ou bien les deux. C'est ainsi qu'à partir de 60 séries originales, nous allons faire notre choix parmi 277 séries.

Les séries mensuelles originales sont de trois types : données quantitatives en volume, données qualitatives (i.e. données d'enquête) et données financières. Les données quantitatives sont *a priori* nécessaires car c'est en priorité celles utilisées par l'INSEE pour la première estimation du PIB, mais comme elles sont coïncidentes, elles doivent le plus souvent être prévues pour permettre une estimation précoce de la croissance, car elles sont connues avec un certain délai. C'est pourquoi nous en retenons un nombre minimum, à savoir 13 : 2 indices de la production industrielle (manufacturière et totale [3]), 9 consommations mensuelles, 1 série d'exportations et 1 série d'importations [4]). Les autres indices de la production ont été exclus car ils n'existent que depuis 1990. Les données mensuelles d'enquêtes sont au nombre de 40 (enquête mensuelle dans l'industrie, le commerce, le bâtiment et les services, enquête auprès des ménages), provenant majoritairement de l'INSEE, complétées par certaines séries de la Commission européenne. Elles sont coïncidentes ou avancées et connues plus rapidement que les données quantitatives. Les données financières sont au nombre de 7 (taux de change réel du dollar, indice de la Bourse de Paris, prix réel du pétrole, variation d'un taux court, variation

1. Préalablement stationnarisées et ensuite standardisées.
2. Bien entendu, on raisonne avec une série dont le niveau est stationnaire.
3. Comprenant les IAA et l'énergie en plus de l'industrie manufacturière.
4. Séries des Douanes, déflatées par nos soins en mensualisant les prix des exportations et des importations issus de la comptabilité trimestrielle.

d'un taux long et écart entre les taux long et court. Ces variables financières sont avancées et connues sans délai. Le marché boursier français, très relié au marché boursier américain, est un bon indicateur de la santé des économies étrangères, non directement représentées ici.

L'algorithme LARS est apparu approprié pour obtenir un premier classement indicatif des 277 séries. Comme les données sont mensuelles et que la cible – le taux de croissance du PIB réel – est trimestrielle, il faut commencer par « trimestrialiser » les 60 séries de la base de données (ceci s'est fait par somme ou moyenne selon le cas) ; ensuite, des transformations diverses sont réalisées (calcul de taux de croissance, de différence première, retards). L'algorithme est mis en œuvre pour des données couvrant la période 1988-2007, le démarrage en 1988 permet d'intégrer l'indicateur résumé de l'enquête dans les services, qui n'existe que depuis cette date.

À titre indicatif, le premier classement fait apparaître dans les 20 premières séries de la sélection LARS : 4 données quantitatives [5] (les 2 indices de la production industrielle, la consommation en produits manufacturés et les exportations), 3 séries financières (l'indice du marché boursier français [6] retardé d'un et de deux trimestres, le prix réel du pétrole en euros [7] retardé de quatre trimestres) et 13 données d'enquête [8] (également réparties entre les enquêtes dans l'industrie, le commerce de détail, le bâtiment et auprès des ménages) dont 7 retardées. Ensuite, nous avons procédé à un deuxième classement excluant les séries financières. En effet, il est apparu préférable de construire des facteurs sans elles, et de les intégrer directement comme régresseurs à côté des facteurs. Ce choix sera justifié ultérieurement.

On s'est alors demandé, pour ce deuxième classement, si la sélection par la méthode LASSO allait donner un classement voisin ou non. Pour éviter de travailler sur un grand nombre de séries, nous avons retenu comme base de départ les 50 premières séries du classement LARS et mis en œuvre le programme d'optimisation quadratique présenté ci-dessus pour la méthode LASSO, en faisant varier le second membre de la contrainte, noté C_1. Si, partant de la borne C_1 correspondant aux 50 variables [9], on fait décroître la valeur de cette borne, alors le nombre de variables de la sélection diminue (un certain nombre de coefficients de régression s'annulent). On s'arrêtera quand on aura atteint le nombre souhaité. En ordonnant les coefficients non nuls par valeur absolue décroissante, on obtient un classement des séries. Un inconvénient majeur de cette méthode est que la décroissance du nombre de série n'est pas monotone et qu'ainsi les différentes sélections, lorsque C_1 croît, ne sont pas nécessairement incluses les unes dans les autres. Par exemple, pour C_1=0,94, la sélection comprend 14 variables correspondant exactement aux 14 premières variables de la sélection LARS ; pour C_1=0,95, on obtient seulement 12 variables, car deux séries sont sorties de la sélection, séries qui ne réapparaîtront plus ensuite ; ce qui fait que pour C_1=1,15, on obtient à nouveau 14 variables mais ce ne sont plus les mêmes que celles de la sélection C_1=0,94 (deux séries distinguent les deux sélections). Cet inconvénient nous fait préférer la méthode LARS puisqu'elle donne une sélection unique. La méthode EN présente le même inconvénient que la méthode LASSO, tout en étant encore moins contrôlable, car elle fait intervenir deux bornes. Elle n'a pas été expérimentée dans ce travail.

Revenons au cas où les méthodes LARS et LASSO donnent exactement la même sélection (tableau 1). La série classée en premier dans les deux cas est la production

5. En volume et en taux de croissance.
6. En taux de croissance. L'indice choisi est l'indice MSCI. Il est déflaté par l'indice des prix à la consommation.
7. En taux de croissance.
8. Aucun des indicateurs synthétiques résumant les enquêtes ne figure parmi ces variables.
9. La borne C_1 est alors la somme des valeurs absolues des coefficients de la régression sur ces variables.

industrielle manufacturière. Les trois autres variables quantitatives retenues apparaissent dans les deux sélections dans les sept premières places : les exportations (respectivement à la 3e et 5e place), la production industrielle totale (respectivement à la 4e et 7e place) et la consommation manufacturière (à la 6e place). Les variables d'enquête les mieux classées viennent de l'enquête industrie et sont en niveau et coïncidentes. Par deux fois, une variable d'enquête est sélectionnée sous deux formes : en niveau et différence première (opinion sur l'évolution de la production prévue), courante et retardée (opinion des ménages sur la possibilité d'épargner). Aucune variable synthétique résumant les enquêtes n'apparaît. L'enquête service est la seule absente.

Tableau 1 : Les quatorze premières séries des classements LARS et LASSO

Rang LARS	Définition des séries et source	Transformation	Avance (en trim.)	Rang LASSO
1	Indice de la production manufacturière (En volume, Insee)	Taux de croissance	0	1
2	Opinion sur l'évolution de la production prévue (*) (Enquête industrie manufacturière, Insee)	Niveau	0	12
3	Exportations (Service des douanes, calcul Ofce pour le volume)	Taux de croissance	0	5
4	Indice de la production industrielle (En volume, Insee)	Taux de croissance	0	7
5	Opinion sur le niveau des stocks (Enquête industrie, Insee)	Niveau	0	3
6	Consommation en produits manufacturés (En volume, Insee)	Taux de croissance	0	6
7	Opinion des ménages sur la possibilité d'épargner (**) (Enquête auprès des ménages, Insee)	Niveau	1	4
8	Opinion sur l'évolution passée de l'activité dans le bâtiment (Enquête industrie du bâtiment, Insee)	Variation	2	8
9	Opinion des ménages sur la possibilité d'épargner (**) (Enquête auprès des ménages, Insee)	Niveau	0	10
10	Opinion sur les effectifs prévus dans le bâtiment (Enquête construction, CE)	Variation	1	14
11	Opinion sur l'évolution de la production prévue (Enquête industrie, Insee)	Niveau	0	2
12	Opinion sur les carnets de commande dans le bâtiment (Enquête dans l'industrie du bâtiment, Insee)	Variation	2	11
13	Opinion sur l'évolution de la production prévue (*) (Enquête industrie manufacturière, Insee)	Variation	0	9
14	Intentions de commandes dans le commerce de détail (Enquête commerce de détail, CE)	Variation	2	13

(*) et (**) variables qui apparaissent sous deux formes.
Source : Calculs de l'auteur.

On constate que ces méthodes n'éliminent pas des variables fortement corrélées entre elles comme par exemple les deux indices de la production industrielle, ou encore, la production industrielle manufacturière et l'opinion sur l'activité dans l'industrie

manufacturière. Cette éventuelle colinéarité entre les séries n'est pas gênante ici puisqu'on s'apprête à extraire les fluctuations communes d'un ensemble de séries.

4. L'information servant au calcul des facteurs

La principale question qui se pose au sujet des facteurs est le nombre de séries à retenir dans la sélection LARS (puisque nous optons pour cette méthode) pour les calculer [10]. Bai et Ng retiennent 30 séries sur 132 (donc un peu plus d'un quart), sans justifier ce choix. Après un long tâtonnement dont nous allons donner les grandes lignes, notre choix s'est porté sur les 14 séries du tableau 1, pour obtenir des facteurs qui, rappelons-le, ne contiennent pas les chocs financiers. Ceux-ci interviendront directement par la suite. Quels sont les problèmes que nous avons rencontrés ?

D'abord, le classement qui apparaît sur la période (1988-2007) n'est pas le même que le premier que l'on aurait dû considérer (période 1988-2000) et que les suivants qui auraient pu être faits en temps réel, puisque notre analyse porte sur la période 2001-2007. Notons que les données quantitatives figurent toujours aux premières places des différents classements, seules les séries d'enquêtes peuvent différer, en particulier dans les tous premiers classements. Sur les cinq dernières années, le classement est beaucoup plus stable. Afin de ne pas compliquer une simulation déjà très lourde, nous conservons la même sélection dans toutes nos expérimentations. On peut penser que si elle avait été modifiée (au moins pour les premiers classements) les résultats présentés se trouveraient améliorés.

Si l'on examine la valeur du critère BIC au fur et à mesure que l'on ajoute les séries de la sélection LARS, il atteint son minimum pour 7 séries. Cependant, on ne va pas choisir ce nombre car alors la sélection comprend 6 séries coïncidentes contre une seule avancée. Or les séries coïncidentes ne sont pas entièrement connues au moment de la prévision et vont donc être une source importante d'erreur. Pour faire une part plus grande aux séries avancées, il faut augmenter le nombre de séries de la sélection pour introduire plus de variables avancées sans trop détériorer l'estimation. C'est ainsi qu'après tâtonnement, on est arrivé à la sélection du tableau 1 pour construire les facteurs. À partir de 14 séries, on détermine 14 composantes principales dont seulement un petit nombre vont figurer comme régresseurs dans l'équation de prévision, celles qui s'avèreront significatives. Cette liste de régresseurs est susceptible de changer au cours du temps lors de l'expérimentation en temps réel. On s'aperçoit que plus les données initiales sont hétérogènes, plus le nombre de facteurs significatifs est grand (ce qui est logique), mais aussi, plus la liste des facteurs significatifs est instable lors de l'expérimentation en temps réel sur la période 2001-2007. On peut ainsi se retrouver à chaque date de prévision avec plusieurs facteurs dont le seuil critique est autour de 5 % et rester alors indécis quant à ceux qui doivent ou non être maintenus. C'est pour cette raison que nous avons choisi d'exclure les séries financières du calcul des facteurs. On peut en effet penser *a priori* qu'à chaque série financière est associée un facteur, car chacune représente un choc spécifique. Dans ce cas, il est préférable d'introduire les séries financières directement plutôt que le facteur qui leur serait associé.

Nous avons mentionné que Bai et Ng s'étaient intéressés à des modèles « non linéaires » en considérant soit des facteurs non linéaires, soit des facteurs linéaires élevés au carré dans l'équation de prévision. Il faut dire que leur prévision porte sur le taux d'inflation où des non linéarités de ce type pourraient intervenir [11]. Pour prévoir la croissance, de telles non

10. Un facteur est une combinaison linéaire des séries de la sélection choisie dont les poids sont donnés par les vecteurs propres de la matrice de corrélation.

linéarités ne s'imposent pas, comme nous avons pu le vérifier après expérimentation. Travailler avec des facteurs non linéaires est particulièrement nocif car ils sont à l'origine d'une très grande instabilité de la liste des facteurs significatifs.

5. Détermination du modèle factoriel

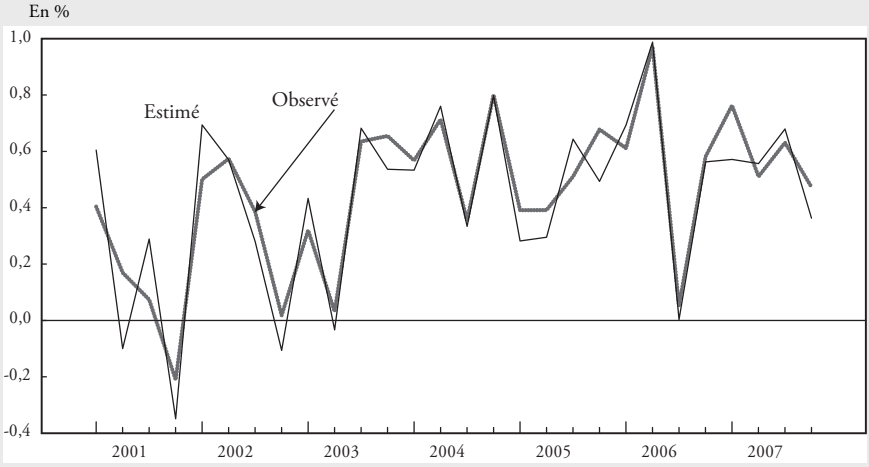
Pour un trimestre T, la première estimation de la croissance trimestrielle du PIB est publiée par l'Insee un mois et demi après la fin du trimestre (respectivement mi-février, mi-mai, mi-août et mi-novembre). On connaît alors toutes les grandeurs quantitatives du tableau 1 pour le trimestre T, ainsi que les données d'enquête. En effet, les indices de la production industrielle du dernier mois (m_3) du trimestre T sont connus quelques jours avant [12], les exportations [13] au début du mois m_3, et la consommation manufacturière [14], environ 3 semaines après le début du mois m_2.

Pour établir les modèles factoriels utilisés pour les estimations de la section suivante, nous nous plaçons juste après la parution des indices de la production industrielle (au plus 5 jours avant la première estimation du PIB par l'Insee), ce qui signifie que toutes les données utilisées sont connues sur le trimestre à estimer. Notre première estimation, celle de la croissance du 1er trimestre 2001, aurait eu lieu aux environs du 10 mai 2001 avec l'information connue à cette date. Ensuite, nous passons à l'estimation de la croissance du 2e trimestre 2001, vers le 10 août 2001, etc., jusqu'à la croissance du 4e trimestre 2007 vers le 10 février 2008. Au final, 28 estimations vont être données. Pour chacune, il faut extraire les composantes principales de la sélection du tableau 1, puis arrêter une régression sur une période qui se termine un trimestre avant celui à prévoir et qui spécifie la relation entre le taux de croissance du PIB et certaines composantes principales, mais aussi des variables financières. Les composantes principales sont numérotées par importance décroissante [15]. Dans les 28 régressions, 2 ou 3 composantes principales (CP) sont significatives. La 1re CP apparaît dans les 28 régressions, la 7e CP dans 21 régressions (elle ne figure pas dans les 7 premières), la 6e CP dans les 9 dernières régressions et, enfin, la 10e CP figure dans les 14 premières régressions. Concernant les variables financières, on a testé la significativité de celles apparues dans la première sélection LARS. Nous avions mentionné que dans les 20 premières séries classées apparaissaient le taux de croissance de l'indice de la bourse de Paris retardés de 1 et de 2 trimestres et le taux de croissance du prix réel du pétrole en euros retardé de 4 trimestres. Au-delà de 20, on trouve la variation du taux d'intérêt à 3 mois retardée de 3 trimestres (au 22e rang) et retardée de 4 trimestres (au 24e rang), l'écart entre le taux long et le taux court retardé de 4 trimestres (au 29e rang) et le taux de croissance du taux de change réel du dollar retardé de 2 trimestres (au 30e rang). Dans les 28 régressions, les variables financières finalement retenues sont la variation du taux de croissance de l'indice boursier [16] (avec un signe positif, i.e. une augmentation du taux de rentabilité du marché boursier laisse présager une augmentation de la croissance), le taux de croissance du

11. Le modèle ARCH de Engle a initialement été développé pour une modélisation du taux d'inflation.
12. Ces indices paraissent environ le 10 de chaque mois.
13. La série en valeur paraît aux alentours du 8 de chaque mois.
14.La consommation manufacturière d'un mois m paraît aux alentours du 23 du mois (m-2)
15. La 1re est associée à la valeur propre la plus élevée de la matrice de corrélations des 14 séries, … , la 14e, à la valeur propre la plus faible de cette matrice.
16. Le taux de croissance du marché boursier retardé d'un trimestre a un coefficient significativement positif, tandis que le taux retardé de deux trimestres, significativement négatif. Les deux coefficients ayant une valeur absolue non significativement différente, c'est la variation du taux qui est introduite dans le modèle.

prix réel du pétrole en euros retardé de 4 trimestres (avec un signe négatif) et le taux de croissance du taux de change réel du dollar retardé de 2 trimestres (avec un signe positif, la hausse du dollar ou la baisse de l'euro favorisera la croissance). Aucune variable de taux d'intérêt n'est retenue. L'écart entre le taux long et court ajouté aux trois variables précédentes (bourse, pétrole et dollar) n'est pas significatif ; il en est de même de la variation du taux court retardée de 4 trimestres. Par contre, la variation du taux court retardée de 3 trimestres est significative mais avec un signe positif, donc incorrect. Les estimations obtenues avec ces régressions sont reportées sur le graphique 1.

Graphique 1 : Taux de croissance trimestriel du PIB observé et estimé lorsque les données mensuelles du trimestre sont connues



Sources : INSEE, calculs OFCE.

On rappelle que pour l'estimation du trimestre T, la régression s'arrête en (T–1), mais les régresseurs sont connus en T. Ces estimations sont sans biais (selon le test spécifié en note 18) et conduisent à une erreur moyenne quadratique de 0,12 point.

Dans un deuxième temps, on a cherché à estimer la croissance sans utiliser de données quantitatives, avec uniquement des données d'enquêtes et des séries financières. Pour cela, les données d'enquêtes ont été classées par la méthode LARS et nous avons retenu les 9 premières séries du classement (tableau 2). Ce nombre a été arrêté après comparaison des sélections comprenant respectivement 7 (nombre qui minimise le critère BIC), 9, 10 et 13 séries. La comparaison est relativement simple car seule la première composante principale sert à expliquer le taux de croissance du PIB (les autres ne sont pas significatives). Les données sont en effet très homogènes et le résumé par un seul facteur est suffisant. Finalement, dans les 28 régressions qui permettront d'estimer la croissance du PIB figurent à côté de la 1re CP les trois variables financières précédemment retenues.

Les estimations obtenues avec ces régressions sont reportées sur le graphique 2. Elles sont sans biais et conduisent à une erreur moyenne quadratique de 0,17 point, significativement supérieure à celle du graphique 1.

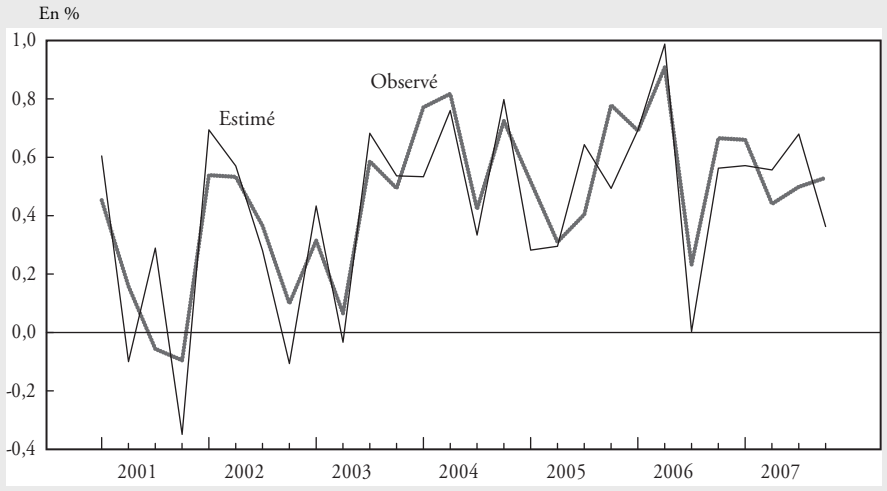
Mais le jeu de données utiles à ces estimations est disponible, non pas 5 jours avant la sortie du PIB, mais 45 jours avant. La comparaison doit donc être faite avec l'estimation de

Tableau 2 : Les neuf premières séries d'enquêtes du classement LARS

Rang LARS	Définition des séries et source	Transformation	Avance (en trim)
1	Opinion sur l'évolution de la production prévue (*) (Enquête industrie manufacturière, Insee)	Niveau	0
2	Opinion sur les effectifs prévus dans le bâtiment (Enquête construction, CE)	Variation	1
3	Opinion sur l'évolution de la production prévue (*) (Enquête industrie, Insee)	Variation	0
4	Indicateur synthétique (Enquête industrie manufacturière, Insee)	Variation	0
5	Opinion sur le niveau des stocks (Enquête industrie manufacturière, Insee)	Niveau	0
6	Opinion sur l'évolution passée de l'activité dans le bâtiment (Enquête industrie du bâtiment, Insee)	Variation	2
7	Opinion des ménages sur la possibilité d'épargner (Enquête auprès des ménages, Insee)	Niveau	1
8	Opinion sur les effectifs prévus dans le bâtiment (Enquête industrie du bâtiment, Insee)	Variation	0
9	Opinion sur les carnets de commande dans le bâtiment (Enquête dans l'industrie du bâtiment, Insee)	Variation	2

Source : Calculs de l'auteur.

Graphique 2 : Taux de croissance trimestriel du PIB observé et estimé lorsque les données d'enquête du trimestre sont connues



Sources : INSEE, calculs OFCE.

la croissance avec une information quantitative à la même date, ce que nous allons faire dans la section suivante.

6. Extrapolation des séries mensuelles et estimations précoces de la croissance

Nous allons considérer l'information à trois dates différentes dans un mois m : (i) vers le 10, soit après la parution des exportations et des indices de la production industrielle (IPI) du mois (m–2), (ii) un peu après le 20, soit après la parution des consommations mensuelles du mois (m–1), (iii) vers le 30, soit après la parution des enquêtes du mois m par l'Insee et par la CE. Dans ce calendrier simplifié, la date de parution du PIB est le 15 du mois et on va examiner les estimations précoces que l'on peut donner 25 jours avant la parution du PIB, 35 jours avant, et de même, de 45 à 165 jours avant, en progressant avec un pas de 10 jours. Ce décalage constant de 10 jours ne correspond pas exactement à la réalité, mais ceci n'est guère important, les dates exactes d'estimation étant le jour de la parution des IPI, celui de la parution de la consommation manufacturière et celui de la parution des enquêtes de la CE [17]. Le tableau 3 indique, pour chaque date, le nombre de mois non connus de chaque série de la sélection des tableaux 1 et 2, et la méthode d'extrapolation choisie pour les prévoir.

Quand on se place de 5 à 85 jours avant la parution du PIB du trimestre T, on estime ce dernier connaissant le trimestre (T–1), au-delà, on estime le trimestre T, connaissant uniquement le trimestre (T–2). Bien que nous considérions que notre procédure ne convient pas pour la prévision à un horizon de deux trimestres, nous avons malgré tout couvert cette situation sans changer de méthode. En effet, si l'on doit faire une prévision de deux trimestres consécutifs, nous pensons qu'il vaut mieux le faire avec la même procédure plutôt que d'adopter des procédures différentes selon l'horizon de prévision. Par exemple, pour un horizon de deux trimestres, on pourrait penser faire une sélection LARS n'utilisant que des séries retardées, puis calculer les facteurs associés et déterminer à nouveau les régressions appropriées. Mais alors les modèles seraient différents de ceux utilisés pour l'horizon 1 et enchaîner des prévisions provenant de modèles distincts peut poser problème, lorsque les prévisions à l'horizon 1 des modèles utilisés pour l'horizon 2 sont très différentes des prévisions à l'horizon 1 des modèles utilisés pour l'horizon 1.

Pour prévoir les données quantitatives, on utilise, en plus des termes autorégressifs, des données d'enquêtes et des données financières. Les équations de prévisions sont spécifiées en taux de croissance mensuel des variables. Pour chacune, un classement des régresseurs éventuels a été établi par la méthode LARS. Ces classements sont basés sur des sélections préliminaires comprenant divers retards de chaque série d'enquête et plusieurs formes (niveau et variation), des termes autorégressifs de la cible ainsi que des séries financières. En effet, ces classements ne vont pas servir maintenant à calculer des facteurs, mais sont utilisés comme un moyen rapide de sélectionner des régresseurs potentiels. Pour chaque cible (les indices de la production industrielle, les exportations et la consommation en produits manufacturés), on a retenu dans chacune des 84 régressions (7 ans × 12 mois) et pour chaque horizon (fonction du nombre de mois à prévoir : 1, 2 ou 3 mois) les variables qui minimisaient le critère BIC, puis nous avons éliminé de chacune, les variables qui n'étaient pas significatives.

Voici brièvement le contenu de ces nombreuses régressions. Pour modéliser la consommation, toutes les équations comportent 2 termes autorégressifs, la variation du taux d'intérêt court retardée de 4 mois, les opinions des ménages sur les intentions d'achat important et sur le niveau de vie futur (en variation et retardées de 1 ou 2 mois selon l'horizon de prévision). Pour modéliser les exportations, toutes les équations contiennent

17. Qui paraissent un peu après celle de l'Insee, mais avant la fin du mois.

Tableau 3 : Nombre de mois à prévoir selon le type de données et la date de prévision

Nombre de jours avant la parutiondu PIB	Nombre de mois à prévoir	Méthode d'extrapolation
25 j	1 mois d'IPI et d'exportations	Régression (*)
35 j	1 mois d'IPI et d'exportations 1 mois de consommation manuf.	Régression Régression
45 j	2 mois d'IPI et d'exportations 1 mois de consommation manuf.	Régression Régression
55 j	2 mois d'IPI et d'exportations 1 mois de consommation manuf. 1 mois de données d'enquêtes	Régression Régression Estimé par la moyenne des 2 mois connus
65 j	2 mois d'IPI et d'exportations 2 mois de consommation manuf. 1 mois de données d'enquêtes	Régression Régression Estimé par la moyenne des 2 mois connus
75 j	3 mois d'IPI et d'exportations 2 mois de consommation manuf. 1 mois de données d'enquêtes	Régression Régression Estimé par la moyenne des 2 mois connus
85 j	3 mois d'IPI et d'exportations 2 mois de consommation manuf. 2 mois de données d'enquêtes	Régression Régression Estimés par le 1er mois du trimestre
95 j	3 mois d'IPI et d'exportations 3 mois de consommation manuf. 2 mois de données d'enquêtes	Régression Régression Estimés par le 1er mois du trimestre
105 j	4 mois d'IPI et d'exportations 3 mois de consommation manuf. 2 mois de données d'enquêtes	Modèle AR(p) Régression Estimés par le 1er mois du trimestre
115 j	4 mois d'IPI et d'exportations 3 mois de consommation manuf. 3 mois de données d'enquêtes	Modèle AR(p) Modèle AR(p) Modèle AR(p)
Au delà	Jusqu'à 6 mois d'IPI et d'export. Jusqu'à 5 mois de conso. manuf. Jusqu'à 5 mois d'enquêtes	Modèle AR(p) Modèle AR(p) Modèle AR(p)

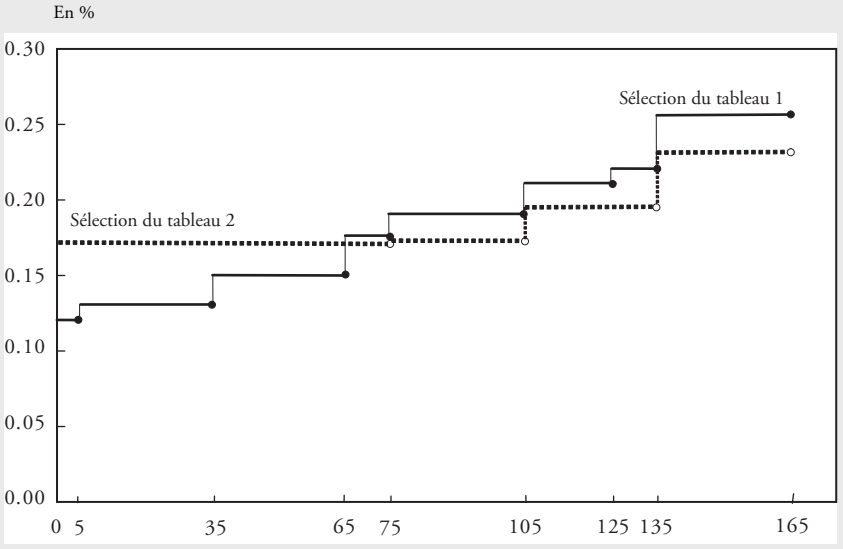
Note : (*) où les régresseurs sont des termes autorégressifs, des données d'enquêtes et éventuellement des séries financières.
Source : Calculs de l'auteur.

2 termes autorégressifs, le taux de croissance de l'indice boursier retardé de trois mois et les variations retardées (1 ou 2 mois) de deux séries de l'enquête dans l'industrie, à savoir les perspectives générales de production et l'indice synthétique. Concernant l'indice de la production industrielle de l'ensemble de l'industrie, les équations de prévision comportent 3 termes autorégressifs et plusieurs soldes d'opinion de l'enquête industrie en variation retardées (production passée et prévue, intentions de commande et niveau des stocks). Les équations donnant l'indice de la production manufacturière diffèrent selon la date, y compris concernant le nombre de termes autorégressifs (1 ou 2). Les soldes d'opinion de l'enquête industrie qui apparaissent, sont le plus souvent, les productions passée et prévue et les intentions de commande, en variations retardées. Lorsque l'horizon de prévision dépasse 3 mois, toutes les grandeurs quantitatives sont extrapolées avec des modèles AR(p). Notons que, pour ces grandeurs mensuelles très volatiles, les modèles de régressions ne sont généralement pas supérieurs aux modèles AR(p) lorsque l'horizon de prévision dépasse un mois.

Concernant les données d'enquête, si le trimestre est incomplet, il est remplacé par la moyenne des deux mois disponibles du trimestre ou par l'unique mois disponible du trimestre. Par contre, si aucune donnée d'enquête n'est disponible sur le trimestre à estimer, on extrapole les données d'enquête à l'aide de modèles AR(p).

De cette manière, j jours avant la parution du PIB du trimestre T, on dispose de données pour le trimestre T (et le trimestre (T–1) s'il n'est pas connu, $j \geq 95$). On calcule alors les facteurs de la sélection du tableau 1 pour les 28 dates de type j. Les régressions établies dans la section précédente servent à nouveau, sur une période se terminant soit en (T–1) soit en (T–2) (selon j) et une estimation de la croissance du trimestre est donnée car les facteurs sont « connus » jusqu'en T. On fait de même avec la sélection qui ne contient que des données qualitatives. Sur le graphique 3, on a reporté, en abscisse, les dates d'estimation exprimées en nombre de jours avant la parution du PIB et, en ordonnée, les erreurs quadratiques moyennes pour les modèles factoriels basés sur les deux sélections des tableaux 1 (trait plein) et 2 (trait pointillé). Nous avons vérifié par un test [18] l'absence de biais de toutes les estimations du graphique 3.

Graphique 3 : Erreurs quadratiques moyennes (en point) sur le taux de croissance trimestriel du PIB (en %) selon la date d'estimation



Source : calculs OFCE.

Sur le graphique 3, le PIB du trimestre T paraît en 0, celui du trimestre (T-1) en 90, et celui du trimestre (T-2) en 180. Voici un exemple de lecture de la fonction en escalier en trait plein du graphique 3. Entre (0 et 5] jours avant la parution du PIB, l'erreur moyenne est de 0,12 point ; de]5 à 35] jours, elle est de 0,13 point ; de]35 à 65] jours, elle est de 0,15 point ; de]65 à 75] jours, elle passe à 0,18 point, et devient supérieure à la fonction en

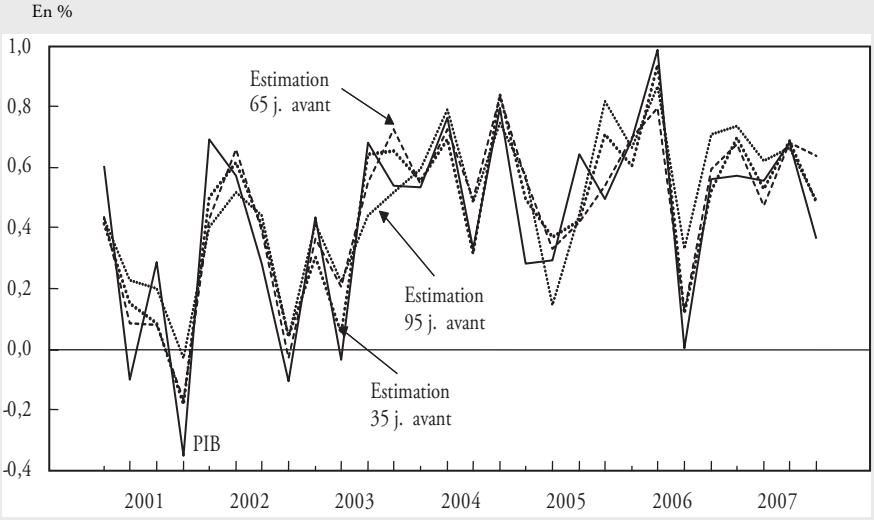
18. Si y est la cible et y^* son estimation, cette dernière est sans biais si on accepte l'hypothèse {a=1 et b=0} dans la régression $y = a\,y^* + b$.

escalier en trait pointillé (0,17 point). Cette dernière n'augmente pratiquement pas avant 105 jours, date après laquelle les données d'enquêtes sur le trimestre à prévoir ne sont plus disponibles.

De deux mois avant la parution du PIB jusqu'à sa parution, l'introduction de données quantitatives permet incontestablement une meilleure estimation. Deux mois et demi avant (abscisse 75), les deux sélections donnent des erreurs moyennes équivalentes. Ensuite, la sélection du tableau 2 conduit en moyenne à moins d'erreur. Finalement, pour une estimation du trimestre T connaissant (T–1), la sélection du tableau 1 est préférable bien qu'il faille extrapoler les données quantitatives non financières. Pour une estimation du trimestre T connaissant (T–2), mieux vaut se passer de ces dernières.

Sur le graphique 4 sont reportées les estimations faites respectivement 35, 65, 95 jours avant la parution du PIB avec la sélection du tableau 1 utilisant des données quantitatives. La courbe représentant les estimations données 95 jours avant la parution du PIB T, soit 5 jours avant la parution du PIB (T–1), est assez différente des deux autres courbes. Il faut dire qu'à cette date, on ne connaît aucune donnée quantitative sur le trimestre à prévoir et seulement 1 mois d'enquêtes.

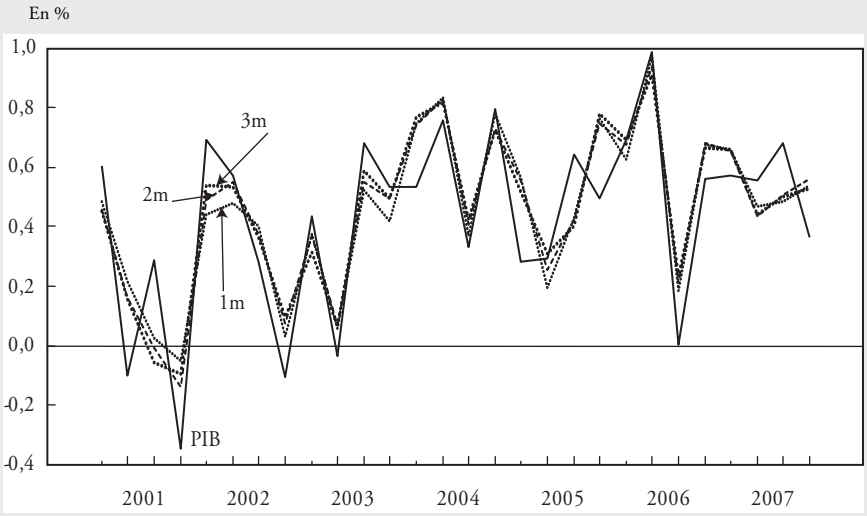
Graphique 4 : Estimations de la croissance trimestrielle respectivement 35, 65 et 95 jours avant sa parution, avec les données du tableau 1



Sources : INSEE, calculs OFCE.

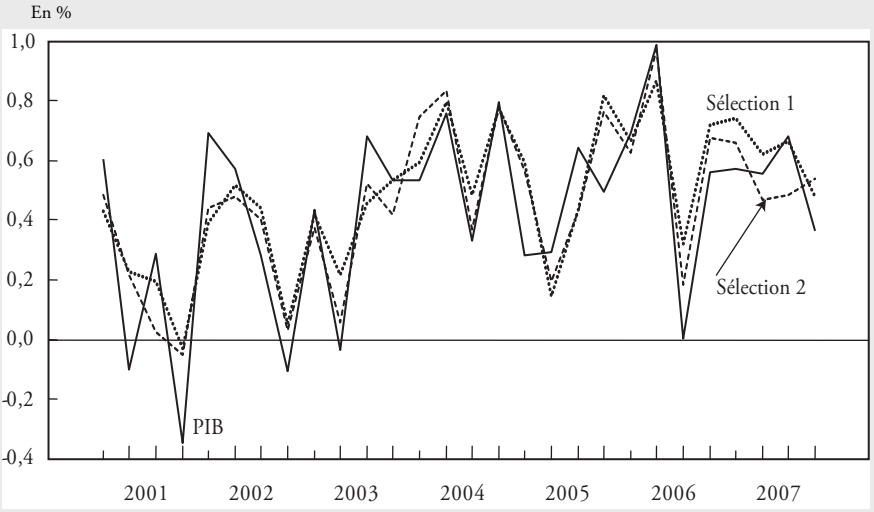
Sur le graphique 5 sont reportées les estimations réalisées avec la sélection du tableau 2 lorsqu'on connaît respectivement les trois mois d'enquêtes du trimestre à prévoir, deux mois d'enquêtes et enfin un seul (soit respectivement 45, 75 et 105 jours avant la parution du PIB). Sur la période (2004-2007) les trois courbes sont extrêmement proches. Ainsi, la manière sommaire adoptée ici pour compléter les mois d'enquêtes manquants ne dégrade

Graphique 5 : Estimations de la croissance trimestrielle connaissant respectivement 3, 2 et 1 mois d'enquête du trimestre estimé, avec les données du tableau 2



Sources : INSEE, calculs OFCE.

Graphique 6 : Estimations de la croissance trimestrielle du PIB (*T*) juste après la parution du PIB (*T*-1) avec les données des tableaux 1 et 2



Sources : INSEE, calculs OFCE.

pas vraiment les estimations. Enfin sur le graphique 6, on compare les estimations du PIB T obtenues avec les deux sélections, juste après la sortie du PIB (T–1), soit 85 jours avant (l'erreur quadratique moyenne de la sélection du tableau 2 est meilleure, inférieure environ de 0,02 point).

Conclusion

L'algorithme de la régression LARS nous est apparu très utile pour obtenir une sélection de séries ciblées parmi toute l'information mensuelle disponible, mais aussi, pour déterminer leur caractère – coïncident ou avancé – ainsi que leur forme – niveau ou variation dans le cas des données qualitatives. Concernant les enquêtes de conjoncture, on a souvent tendance à privilégier les indices résumés, alors qu'ici ces résumés ne sont pas retenus. La méthode LARS permet justement de choisir rapidement entre tous les soldes d'opinion disponibles et, même, entre divers fournisseurs de données d'enquêtes.

Le passage par les facteurs permet d'introduire toutes les séries de la sélection, ce qui ne serait pas possible sinon, vu la colinéarité existant entre les séries. L'estimation est ainsi basée sur une information un peu plus large et devient moins dépendante du comportement de chaque série individuelle.

La méthode décrite ici est simple à mettre en œuvre et donne des résultats *a priori* satisfaisants pour le trimestre coïncident. Il faut tout de même rappeler que les résultats des graphiques 3, 4, 5 et 6 proviennent d'une pseudo analyse en temps réel, et seraient probablement nettement moins bons s'il s'agissait d'une véritable analyse en temps réel.

Références bibliographiques

Bai J. et S. Ng, 2002, « Determining the Number of Factors in Approximate Factor Models », *Econometrica,* 70:1, 191-221.

Bai J., 2003, « Inference on Factor Models of Large Dimension », *Econometrica,* 71:1, 135-172.

Bai J. et S. Ng, 2006, « Confidence Intervals for Diffusion Index Forecasts and Inference with Factor-Augmented Regressions », *Econometrica,* 74:4, 1133-1150.

Bai J. et S. Ng, 2008, « Forecasting Economic Time Series Using Targeted Predictors », *Journal of Econometrics,* 146, 304-317.

Bair E., T. Hastie, D. Paul et R. Tibshirani, 2006, « Prediction By Supervised Principal Components », *Journal of the American Statistical Association,* 101:473, 119-137.

Boivin J. et S. Ng, 2006, « Are More Data Always Better for Factor Analysis », *Journal of Econometrics,* 132, 169-194.

Doz C., D. Giannone et L. Reichlin, 2006, « A quasi maximum likelihood approach for large approximate dynamic factor models », *WP ECB*, n° 674.

Efron B., T. Hastie, I. Johnstone et R. Tibshirani, 2004, « Least Angle Regression », *Annals of Statistics,* 32:2, 407-499.

Giannone D., L. Reichlin, et D. Small, 2006, « Nowcasting GDP and Inflation: The Real Time Informational Content of Macroeconomic Data Releases », *WP ECB*, n° 633, *Journal of Monetary Economics* (forthcoming).

Stock J. H. et M. W. Watson, 1989, « New indexes of coincident and leading economic indicators », *NBER Macroeconomics Annual*, 351-393.

Stock J. H. et M. W. Watson, 2002, « Diffusion Indexes », *Journal of the American Statistical Association,* 97:460, 1167-1179.

Stock J. H. et M. W. Watson, 2002, « Macroeconomic Forecasting Using Diffusion Indexes », *Journal of Business and Economic Statistics,* 20:2, 147-162.

Tibshirani R., 1996, « Regression Shrinkage and Selection via the Lasso », *Journal of Royal Statistical Society,* Series B 58:1, 267-288.

Zou H. et T. Hastie, 2005, « Regularization and Variable Selection via the Elastic Net », *Journal of Royal Statistical Society*, Series B 67:2, 301-320.