Université de Montréal

Prévisions non linéaires pour les modèles à facteurs

par

Joël Corbin-Charland

Département de sciences économiques

Faculté des arts et sciences

Rapport de recherche présenté à la Faculté des arts et des sciences

en vue de l'obtention d'une maîtrise (M.Sc.)

en sciences économiques

29 août 2011

Table des matières

Prévisions non linéaires pour les modèles à facteurs

29 août 2011

Résumé

Ce rapport étudie les prévisions non linéaires pour les modèles à facteurs. Il s'agit essentiellement d'étendre les prévisions du modèle considéré par Stock et Watson dans leur article *Macroeconomic Forecasting Using Diffusion Indexes* par des méthodes non linéaires. Deux méthodes non linéaires ont été sélectionnées ; il s'agit des réseaux de neurones et des séries de Fourier. Nous avons effectué les régressions directement et nous les avons évaluées à l'aide de l'erreur quadratique moyenne de prévision. Les résultats suggèrent qu'il n'y a pas d'avantage à considérer ces méthodes non linéaires dans le cadre des modèles à facteurs. Même en considérant l'incertitude associée à l'évaluation de coefficients supplémentaires pour les modèles non linéaires, nous ne sommes pas en mesure de dire que les modèles non linéaires réduisent l'erreur quadratique moyenne de prévision.

Mots-clés : prévisions, modèles à facteurs, indices de diffusion, non linéaire.

1 Introduction

1.1 Survol

Dans la tradition macroéconométrique, il est d'usage de choisir quelques séries temporelles que l'on trouve pertinentes pour prédire un agrégat. Une difficulté surgit lorsqu'on doit discriminer et se limiter à n'en choisir que certaines. On se base alors sur l'expérience et la théorie économique pour choisir le modèle le plus approprié. Cette manière de prédire a été abondamment étudiée et est celle

qu'on emprunte naturellement lorsqu'on cherche à faire des prévisions. Elle est certainement raisonnable, mais il persiste toujours la possibilité d'oublier une partie substantielle de l'information pertinente. Comme on ne peut pas inclure énormément de variables explicatives dans une régression, il faut procéder à une sélection qui est inéluctablement sujette à la critique.

Par ailleurs, de récents développements concernant les modèles à facteurs à large dimension nous suggèrent une autre approche pour réaliser des prédictions. Cela consiste à condenser l'information de beaucoup de séries en quelques facteurs que l'on nomme indices de diffusion et qui permettent d'effectuer des prévisions. Dès lors que l'on obtient les indices de diffusion, nous sommes amenés à les insérer dans un modèle ; ce dernier pouvant être linéaire ou non linéaire. Dans la mesure où le modèle sous-jacent est linéaire, il convient tout à fait de rester dans le cadre linéaire. Toutefois, cette approche perd sa validité lorsque l'on est en présence de non-linéarité. Ainsi, quand le modèle duquel on extrait les facteurs a une structure non linéaire, il apparaît impératif de recourir à un modèle non linéaire pour établir des prévisions. On notera par ailleurs qu'un modèle non linéaire appliqué à une structure linéaire demeure valide bien qu'il ne soit pas optimal. En étudiant les modèles non linéaires pour les indices de diffusion, on sera en mesure de déterminer si le modèle sous-jacent comporte des non-linéarités. En somme, on cherche à savoir si les prévisions peuvent être rendues plus précises si on considérait les modèles non linéaires.

1.2 Revue de littérature

Les modèles à facteurs ont fait époque en macroéconomie. Leur utilité s'est avérée dans les travaux de Burns et Mitchell (1947) qui y trouvent une pertinence pour les cycles économiques. Plus spécifiquement, on peut distinguer deux principales branches dans le développement des modèles à facteurs. La première se rapporte aux modèles à facteurs dynamiques. Cela consiste à estimer des facteurs non observables à l'aide de l'estimateur du maximum de vraisemblance, du filtre de Kalman ou des deux. On en trouvera des exemples parmi Sargent et Sims (1977), Geweke et Singleton (1981), Engle et Watson (1981), Stock et Watson (1989, 1991) ainsi que Kim et Nelson (1998). La deuxième porte sur les modèles avec indices de diffusion et recourt à la méthode des composantes principales pour évaluer les facteurs communs. Bien que l'estimateur du maximum de vraisemblance est plus efficace pour un petit nombre de séries, cette dernière méthode est plus simple à calculer et c'est pourquoi nous allons nous inscrire dans cette seconde tradition. On pourra se référer à Geweke et Zhou (1996), Sargent (1989), Forni et al. (1998, 2000) ainsi que Stock et Watson (1998, 2002) pour des exemples. Enfin, on notera que la méthode des composantes principales est une approximation, mais que la différence entre ces deux approches s'oblitère à mesure que le nombre de séries à notre disposition s'accroît.

Dans tous les cas, diverses applications des modèles à facteurs suggèrent qu'un petit nombre de facteurs peuvent rendre compte des variations de variables

macroéconomiques. L'utilité avérée de ce paradigme nous incite à vouloir pousser plus avant cette approche. On notera par ailleurs qu'il s'agit d'analyse factorielle exploratoire plutôt que confirmatoire en ce que la première ne cherche qu'à déterminer la structure sous-jacente des facteurs alors que la seconde entend plutôt confirmer une théorie établie *a priori*.

Pour nous situer plus concrètement dans cette littérature, nous allons exposer des résultats sur lesquels nous nous basons pour notre recherche. Stock et Watson (2002b) ont montré que les modèles à facteurs avec beaucoup de prédicteurs (au nombre de N) pour les séries temporelles (de taille T) ont de bonnes propriétés asymptotiques. Les prévisions convergent en probabilité quand $N,\ T \to \infty$ vers la prévision optimale, c'est-à-dire la valeur que l'on obtiendrait si les coefficients étaient connus. Une étude empirique de ces mêmes auteurs (2002a) suggère qu'il serait avantageux d'utiliser des facteurs pour faire des prédictions. Il s'avère en effet que les modèles utilisant les indices de diffusion sont supérieurs à des modèles simples traditionnels comme des VAR.

Toutefois, ces auteurs n'ont considéré que les modèles linéaires. Ainsi, Shintani (2005) s'est penché sur les méthodes non linéaires dans le cas du Japon. Plus spécifiquement, il découvre des non-linéarités dans les données et trouve à nouveau que les modèles utilisant les indices de diffusion performent mieux que les modèles traditionnels. Par contre, la méthode non linéaire employée se basant sur les réseaux de neurones apporte peu d'avantages par rapport aux méthodes linéaires.

Par ailleurs, Bierens, Castelar et Ferreira se proposent d'étendre le modèle de Stock et Watson. Ils tentent de considérer la non-linéarité en permettant la variation temporelle des paramètres. Plus spécialement, ils trouvent que les modèles à changements de régimes markoviens [1] ou avec seuil [2] améliorent les prévisions pour la croissance économique du Brésil. Ces auteurs limitent toutefois leur étude à ce seul agrégat.

Il est donc proposé de refaire l'étude pratique de Stock et Watson (2002a) et d'y développer des prévisions non linéaires. Nos choix, qui ne s'effectuent pas sans réserve, se sont portés sur les réseaux de neurones et sur les séries de Fourier. Pour ce qui est des réseaux neuronaux, une synthèse éclairante fut rédigée par Franses et van Dijk (2000). Cette préférence n'est pas arbitraire étant donné que les réseaux de neurones ont montré de bonnes capacités prédictives dans plusieurs cas. À ce sujet, voir Swanson et White (1997), Chen, Racine et Swanson (2001) ainsi que Hong et Lee (2003). Il se trouve que cette méthode non linéaire est très efficace pour épouser des données. En ce qui a trait aux séries de Fourier, nous reprenons une formule employée par Linton et Perron (2003) qui elle-même est une adaptation de Gallant (1981).

1. Connus plus familièrement en anglais sous le nom de *Markov-switching models*.
2. Traduction de *Treshold model*.

1.3 Données

Il importe de souligner que cette technique est essentiellement économétrique. Elle ne requiert que peu de théorie économique si ce n'est pour la constitution de la base de données, c'est-à-dire la question de savoir quelles séries nous allons inclure comme prédicteurs. En effet, la théorie économique et l'expérience pratique en économétrie permettront de sélectionner les séries jugées les plus significatives pour décrire l'activité économique. Dans l'optique où l'on cherche à constituer une base de données substantielle, il nous incombe de prendre une pléthore de séries, ce qui est sans contredit moins restrictif que lorsqu'on ne choisit que quelques séries comme dans les modèles traditionnels.

En outre, les données sont celles utilisées par Stock et Watson (2002a). Nous disposons donc de 215 séries mensuelles correspondant à des variables macroéconomiques. La période considérée s'échelonne de janvier 1959 à décembre 1998. Toutefois, certaines des séries ne sont pas complètes. On ne prendra que celles qui sont complètes et qui sont au nombre de 146. Dès lors, nous n'utiliserons que le panel cylindrique. De surcroît, des transformations sont appliquées à chacune des séries pour s'assurer de leur stationnarité, c'est-à-dire qu'elles sont $I(0)$. Bien qu'il n'y ait pas de certitude à ce chapitre, différents tests ont été appliqués par Stock et Watson pour s'assurer qu'elles l'étaient effectivement. Puis, l'expérience et le jugement ont servi de dernier rempart contre les méprises. Nous pouvons donc être confiants quant à la validité de ces données. Nous avons ensuite standardisé les séries de telle sorte qu'elles aient une moyenne nulle ainsi qu'une variance unitaire. Enfin, on mentionnera que les valeurs aberrantes [3] ont été remplacées par des valeurs manquantes.

Dans un autre ordre d'idées, on notera que cette base de données n'est pas à strictement parler en temps réel. Cela veut dire que certaines séries ont été révisées au fil du temps et qu'elles incorporent de l'information future. Définissons Ω_t comme l'ensemble d'information dont on dispose à la période t et x_t comme une série jusqu'à ce moment. Au risque d'énoncer un truisme, nous avons que $x_t \in \Omega_t$. Cependant, il arrive que la série x_t soit révisée en ce sens que des observations sont retouchées en regard de l'information subséquente. Dès lors, nous aurons que $x_t \in \Omega_{t+j}$ où $j \in \mathbb{N}$. Cela est problématique dans la mesure où l'on cherche à faire des prédictions seulement à partir de l'information connue à t. Par ailleurs, il se trouve que les données non révisées sont nettement plus difficiles à obtenir. Ainsi, la raison pour laquelle une base de données en temps réel n'est pas utilisée est qu'il aurait été ardu de recueillir la myriade de séries constituant la base de données.

3. On entend par valeur aberrante toute entrée qui était éloignée de la médiane de plus de 10 fois l'écart interquatile.

2 Modèles économétriques

2.1 Modèles à facteurs

Comme nous l'avons mentionné précédemment, nous allons calculer les facteurs par la méthode des composantes principales. Nous avons à notre disposition une grande quantité de séries, au nombre de N, que l'on notera génériquement par $x_{t,\,i}$ avec $i = 1\,,\ldots,\ N$ et $t = 1\,,\ldots,\ T$. L'objectif consiste à prédire le mieux possible une variable scalaire y_{t+h}, où h est l'horizon temporel de notre prédiction, compte tenu de l'information disponible à la période t, Ω_t. Bien entendu, nous ne pouvons pas inclure toutes les séries dans la régression. À la place de ces N séries, nous cherchons r facteurs, c'est-à-dire $f_t = (f_{1,\,t}\,,\ f_{2,\,t}\,,\ldots,\ f_{r,\,t})$, qui nous serviront à établir la prévision. L'intérêt de cette approche consiste à ce que $r \ll N$, c'est-à-dire que r est nettement plus petit que N. Il devient alors pensable de faire une régression avec ces facteurs. On peut ensuite poser plus formellement le modèle.

$$y_{t+h} = \gamma(L)\, y_t + \beta'(L)\, f_t + \epsilon_{t+h}$$

$$x_{t,\,i} = \lambda_i'(L)\, f_t + e_{t,\,i}$$

Où $\gamma(L)$, $\beta(L)$ et $\lambda_i(L)$ sont les coefficients avec les retards pour $i = 1\,,\ldots,\ N$. On prendra note que $\lambda_i(L)$ correspond à la sensibilité de $x_{t,\,i}$ au facteur f_t. Par conséquent, $e_{t,\,i}$ est l'erreur idiosyncrasique propre à la variable $x_{t,\,i}$. On aura également que $E(\epsilon_{t+h} \mid f_t\,,\ y_t\,,\ x_t\,,\ f_{t-1}\,,\ y_{t-1}\,,\ x_{t-1}\,,\ldots) = 0$. Nous allons supposer que le nombre de retards de β et γ est fini de sorte que l'on peut estimer les facteurs par la méthode des composantes principales bien que cela corresponde à une approximation. Cette méthode consiste à trouver λ et f_t qui minimisent $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{t,\,i} - \lambda_i^{(r)\prime}\, f_t^{(r)})^2$ pour un r donné.

Pour décrire plus formellement comment nous allons calculer ces facteurs, nous allons mettre les N séries $x_{t,\,i}$ dans une matrice $X_{T\times N}$. On cherche à en extraire r indices de diffusion notés $F_{T\times r}$. La forme matricielle pour écrire cela est $X_{T\times N} = F_{T\times r}\ \Lambda'_{r\times N} + e_{T\times N}$. La méthode des composantes principales nous donne $\hat{\Lambda}$ comme $\sqrt{N}$ fois les r vecteurs associés aux r plus grandes valeurs propres de la matrice $X'X$. En se servant de la normalisation $\frac{\Lambda'\Lambda}{N} = I$, on obtient que $\hat{F} = \frac{X\hat{\Lambda}}{N}$.

Stock et Watson (2002a) se sont servis de ces indices de diffusion pour faire leurs prédictions. Pour différents agrégats, ils ont considéré le modèle

$$\hat{y}_{t+h|t} = \hat{\alpha} + \sum_{j=1}^{p} \hat{\gamma}_j \, y_{t-j+1} + \sum_{j=1}^{m} \hat{\beta}'_j \, \hat{F}_{t-j+1}$$

On pourra remarquer que cette équation comporte $p + 1 + m \cdot r$ coefficients. Le modèle a été sélectionné en utilisant le critère d'information de Bayes (BIC) avec $1 \leqslant r \leqslant 6, 1 \leqslant m \leqslant 3$ et $0 \leqslant p \leqslant 6$ [4]. Cela fait 126 modèles à estimer pour chaque période. Il importe également de mentionner que les coefficients sont obtenus par les moindres carrés ordinaires puisque les variables entrent linéairement dans l'équation. Cette manière d'estimer des coefficients étant simple, il se trouve qu'elle est rapide à exécuter.

L'essence de la problématique de ce rapport apparaît alors plus nettement. On se demande s'il serait possible d'améliorer cette prédiction en considérant une fonction non linéaire de ces prédicteurs. Plus spécifiquement, on voudrait obtenir une fonction $f : \mathbb{R}^{p+m} \to \mathbb{R}$

$$\hat{y}_{t+h|t} = \hat{\alpha} + \sum_{j=1}^{p} \hat{\gamma}_j \, y_{t-j+1} + \sum_{j=1}^{m} \hat{\beta}'_j \, \hat{F}_{t-j+1} + f\left(y_t \, , \ldots , \; y_{t-p+1} \, , \; \hat{F}_t , \ldots , \; \hat{F}_{t-m+1}\right)$$

telle que f est non linéaire. Il existe de nombreuses méthodes qui permettent d'estimer la partie non linéaire de cette équation. Nous allons nous contenter d'en étudier deux.

2.2 Réseaux de neurones

Avec ces indices de diffusion, nous sommes à même de faire les prévisions à l'aide d'un réseau de neurones. La formule appropriée qui exprime ce qu'est un réseau est la suivante :

$$y_{t+h} = \phi_0 + \phi_1 \, y_t + \ldots + \phi_p \, y_{t-p+1} + \psi'_1 \, \hat{F}_t + \ldots + \psi'_m \, \hat{F}_{t-m+1} +$$
$$\sum_{j=1}^{q} \beta_j \, G\left(\gamma_{0,j} + \gamma_{1,j} \, y_t + \ldots + \gamma_{p,j} \, y_{t-p+1} + \delta'_{1,j} \, \hat{F}_t + \ldots + \delta'_{m,j} \, \hat{F}_{t-m+1}\right) + \varepsilon_{t+h}$$

où $\psi'_i = (\psi_{i,1} \, , \; \psi_{i,2} \, , \ldots , \; \psi_{i,r})$ correspond aux coefficients pour les r facteurs à la période $t - i + 1$. Il en va de même pour $\delta'_{i,j} = \left(\delta^{(1)}_{i,j} \, , \; \delta^{(2)}_{i,j} \, , \ldots , \; \delta^{(r)}_{i,j}\right)$.

4. On aura pour convention que $p = 0$ correspond au fait d'exclure y_t de la régression. On conservera cette convention pour les autres modèles.

De manière condensée, on écrira

$$y_{t+h} = y_{t+h|t} + \varepsilon_{t+h} = x_t' \, \phi + z_t' \, \psi + \sum_{j=1}^{q} \beta_j \, G \left(x_t' \, \gamma_j + z_t' \, \delta_j \right) + \varepsilon_{t+h}$$

avec $\phi = (\phi_0 \, , \ \phi_1 \, , \ldots , \ \phi_p)'$, $\psi = (\psi_1' \, , \ldots , \ \psi_m')'$, $\gamma_j = (\gamma_{0,j} \, , \ \gamma_{1,j} \, , \ldots , \ \gamma_{p,j})'$ et $\delta = (\delta_{1,j}' \, , \ldots , \ \delta_{m,j}')'$ pour les coefficients et $x_t' = (1 \, , \ y_t \, , \ldots , \ y_{t-p+1})$ et $z_t' = \left(\hat{F}_t \, , \ldots , \ \hat{F}_{t-m+1} \right)$ pour les variables. Comme il est coutume de faire, on prend la fonction logistique $G(x) = \frac{1}{1+e^{-x}}$. On remarque que $G : \mathbb{R} \to] \, 0, \ 1 \, [$ et que $G'(x) > 0$. La partie non linéaire est celle constituée par la somme. On peut aussi voir que ce sont les mêmes prédicteurs dans les parties linéaire et non linéaire, mais qu'ils ont des coefficients différents.

On cherche maintenant à évaluer le modèle. Tout d'abord, on définit le vecteur θ qui est composé de tous les $(p + m \cdot r + 1) + q \, (p + m \cdot r + 2)$ paramètres et peut être écrit comme $\theta' = (\phi' \, , \ \psi' \, , \ \beta' \, , \ \gamma' \, , \delta')$ où $\phi_{(p+1)\times 1}$, $\psi_{m \cdot r \times 1}$, $\beta_{q \times 1}$, $\gamma_{q \, (p+1) \times 1}$ et $\delta_{q \cdot m \cdot r \times 1}$. On veut minimiser une certaine fonction d'erreur quadratique à laquelle on adjoindra des poids dégénérés [5] qui empêchent que les coefficients deviennent arbitrairement grands. Ainsi, ils nous permettent d'éviter qu'un élément ait une influence trop importante et ils font en sorte qu'on sélectionne un modèle plus parcimonieux [6]. On définit la fonction à minimiser comme

$$Q(\theta) = \sum_{t=1}^{T-h} (y_{t+h} - y_{t+h|t})^2 + WD(\theta)$$

$$WD(\theta) = \omega_{\phi,\psi} \left(\sum_{i=0}^{p} \phi_i^2 + \sum_{i=1}^{r} \psi_i^2 \right) + \omega_\beta \sum_{j=1}^{q} \beta_j^2 + \omega_{\gamma,\delta} \sum_{j=1}^{q} \left(\sum_{i=0}^{p} \gamma_{i,j}^2 + \sum_{i=1}^{r} \delta_{i,j}^2 \right)$$

La valeur des paramètres sont $\omega_{\phi,\psi} = 0,01$ et $\omega_\beta = \omega_{\gamma,\delta} = 0,0001$, tel que suggéré par Franses et van Dijk (2000). Il importe de sélectionner ces poids à l'avance et de ne pas les définir pour obtenir un certain modèle. À titre indicatif, une reparamétrisation est nécessaire pour que les valeurs des coefficients des poids dégénérés aient un sens. Celle qui a été retenue dicte que $u_t^* = \frac{u_t - min(u_t)}{max(u_t) - min(u_t)}$, ce qui nous assure que toutes les observations sont comprises dans l'intervalle $[\, 0, \ 1 \,]$ [7].

5. Traduction libre de *weight decay*.

6. Un danger avec les réseaux de neurones est de calquer de trop près les données et d'ainsi modeler le bruit, ce qui nuit au pouvoir prédictif.

7. Il aurait également été possible de redéfinir les données comme $u_t^* = \frac{u_t - \bar{u}_t}{\sigma(u_t)}$ où $\bar{u}_t$ est la moyenne empirique et $\sigma(u_t)$ est l'écart-type de la série.

Malheureusement, comme la fonction que l'on souhaite minimiser n'est pas linéaire, il faut recourir à des algorithmes pour résoudre cette équation. Ainsi, les coefficients sont estimés en minimisant $Q(\theta)$ à l'aide de deux algorithmes. Dans un premier temps, on recourt à l'algorithme simplex de Nelder et Mead (1965) qui nous permet de couvrir un plus grand espace que le second. Nous ne l'utilisons qu'une fois afin de préciser notre approximation. Quant au second algorithme, il s'agit de celui de Broyden-Fletcher-Goldfarb-Given (BFGS) qui converge plus rapidement [8]. Il serait opportun de préciser que le temps de calcul pour chacun de ces algorithmes est passablement long, surtout pour le deuxième. Par conséquent, notre approximation risque d'être quelque peu grossière, car nous n'utiliserons le second algorithme qu'au maximum 400 fois. De plus, nous allons chercher les coefficients seulement autour de zéro. L'idéal aurait été de prendre quelques points aléatoirement dans la région où est susceptible d'être le minimum de sorte qu'en trouvant plusieurs minimums locaux, nous aurions plus de chance de trouver le minimum global. Aurions-nous eu un ordinateur plus puissant que nous aurions pu être plus précis, mais faute de moyens, nous nous accommoderons de cette situation.

Le temps de calcul indûment long nous mène à une autre contrainte en ce qui a trait à la sélection du modèle. Comme il devient prohibitif de comparer plusieurs modèles afin de choisir le meilleur, on se contentera de n'en choisir qu'un. Idéalement, nous aurions utilisé le critère du BIC défini ainsi :

$$BIC = (T-h) \cdot ln\left(\frac{1}{T-h}\sum_{t=1}^{T-h}\left(y_{t+h} - \hat{y}_{t+h|t}\right)^2\right) + (p+1+q\,(p+2))\,\frac{ln(T-h)}{T-h}$$

On aurait évalué les modèles pour $1 \leqslant r \leqslant 6$, $0 \leqslant p \leqslant 6$, $1 \leqslant m \leqslant 3$ et $1 \leqslant q \leqslant 5$. Cela fait que pour chaque prévision, on aurait estimé 630 modèles. On aurait pu imaginer une manière de réduire le nombre de modèles à évaluer en fixant certains paramètres. Shintani (2005), par exemple, utilise invariablement six facteurs, car ceux-ci sont supposés bien résumer les séries. Il se trouve que cette modification est insuffisante pour réduire suffisamment le temps de calcul. Nous avons donc opté pour abroger tout simplement la sélection du modèle et nous avons fixé les paramètres à $r = 4$, $p = 3$, $m = 2$ et $q = 2$. Enfin, pour dernière limitation, on mentionnera que l'on ne calcule les coefficients qu'annuellement bien que les données soient mensuelles.

2.3 Séries de Fourier

La seconde approche consiste à utiliser les séries de Fourier. Ainsi, en plus des réseaux de neurones, on emploie une seconde approche pour approximer la partie non linéaire. Elle s'avère pertinente pour tenter de reproduire les cycles dans les données simplement à cause de la forme cyclique de la formule sinusoïdale.

8. On peut voir la procédure détaillée pour implémenter ce dernier algorithme en annexe.

$$y_{t+h} = y_{t+h|t} + \varepsilon_{t+h} = \alpha_0 + \alpha_1\, y_t + \ldots + \alpha_p\, y_{t-p+1} + \beta_1'\, \hat{F}_t + \ldots + \beta_m'\, \hat{F}_{t-m+1}$$
$$+ \sum_{j=1}^{q} \sum_{i=1}^{p} \left(\gamma_{j,i}\, sin\left(j \cdot y_{t-i+1}\right) + \delta_{j,i}\, cos\left(j \cdot y_{t-i+1}\right)\right)$$
$$+ \sum_{j=1}^{q} \sum_{k=1}^{m} \sum_{\ell=1}^{r} \left(\eta_{j,k}^{(\ell)}\, sin\left(k \cdot \hat{F}_{t-k+1}^{(\ell)}\right) + \zeta_{j,k}^{(\ell)}\, cos\left(k \cdot \hat{F}_{t-k+1}^{(\ell)}\right)\right) + \varepsilon_{t+h}$$

On peut voir que ce sont les deux dernières lignes de cette équation qui constituent la partie non linéaire ; la première n'étant que celle déjà utilisée par Stock et Watson. De plus, on remarquera que cette formule comporte $p + 1 + m \cdot r + 2\, q\,(p + m \cdot r)$ coefficients. À chaque itération, on utilisera le BIC pour déterminer le modèle le meilleur modèle à partir des paramètres $1 \leqslant r \leqslant 6$, $0 \leqslant p \leqslant 5$, $1 \leqslant m \leqslant 3$ et $1 \leqslant q \leqslant 4$. Cela fait en tout 432 modèles à estimer à chaque période, ce qui est considérable. On peut se permettre cette prodigalité, car cette équation s'avère nettement plus facile à implanter que pour les réseaux neuronaux. En effet, comme tous les paramètres entrent linéairement dans l'équation, nous estimons ceux-ci par la méthode des moindres carrés ordinaires. Cette dernière étant très rapide à exécuter, cela nous permet d'évaluer le modèle à plusieurs reprises et d'effectuer une sélection à l'aide du BIC. De plus, on peut actualiser les coefficients chaque mois contrairement aux réseaux de neurones où on ne le faisait qu'à chaque année. Enfin, on notera qu'on modifie les données dans ce cas-ci également. On prendra $u_t^* = 2\,\pi\, \frac{u_t - u_{t,\,min}}{u_{t,\,max} - u_{t,\,min}}$ où $u_{t,\,min}$ est plus petit que $min(u_t)$ et $u_{t,\,max}$ est plus grand que $max(u_t)$ [9].

2.4 Prévisions

Il existe plusieurs manières de faire des prévisions. Supposons que nous avons $T + h$ observations et que nous divisons celles-ci en une première partie R et en une autre qui contient toutes les données subséquentes qu'on appellera P. Nous avons que $R + P = T + h$. À partir de ces définitions, il est possible d'estimer les paramètres de trois manières afin de faire des prévisions. La première consiste à toujours utiliser un échantillon de taille R. Pour obtenir la prévision $R + h$, on évalue le modèle avec les données de 1 à R. Pour $R + h + 1$, on prend de 2 à $R + 1$ et ainsi de suite. Cela fait qu'on a une fenêtre qui conserve toujours la même taille à mesure qu'on progresse dans l'échantillon. Cette méthode a l'avantage de nous protéger contre les dérives difficilement modélisables que

9. Plus spécifiquement, on aura pris $u_{t,\,min} = \begin{cases} 1{,}01\, min(u_t) & si\ min(u_t) < 0 \\ 0{,}99\, min(u_t) & si\ min(u_t) \geqslant 0 \end{cases}$ et $u_{t,\,max} = \begin{cases} 1{,}01\, max(u_t) & si\ max(u_t) \geqslant 0 \\ 0{,}99\, max(u_t) & si\ max(u_t) < 0 \end{cases}$

pourraient comporter les paramètres. La seconde approche consiste à n'estimer les paramètres qu'une fois avec l'échantillon allant jusqu'à la R^e observation. Cela a l'avantage de réduire considérablement le temps de calcul, mais néglige une partie significative des données à mesure que l'on approche de la fin de la série à prédire. La troisième méthode consiste à utiliser toutes les données disponibles. Par exemple, pour prédire $t+h$, on utilise les données de 1 à t. Cette manière de procéder est la seule qui emploie toute l'information disponible. Ainsi, seule cette dernière manière a été retenue. Enfin, dans notre cas empirique, nous avons choisi $R = 300$. Considérant que nous disposons d'un échantillon d'environ 480 observations, cela autorise tout de même 180 prévisions. Ce choix peut certes faire l'objet de critique dans la mesure où il n'y a pas d'arguments qui permettent de trancher définitivement la question. Pour leur part, Stock et Watson (2002a) ont choisi 132, ce qui est nettement moindre que dans notre cas. Toutefois, nos modèles comportent plus de coefficients de sorte qu'on perd davantage de degrés de liberté, ce qui justifie que l'on commence nos prévisions plus tardivement. Ce choix s'est aussi effectué en considérant le fait que Shintani (2005) ait sacrifié 64% de son échantillon alors que nous utilisons 63% (300/480) avant de faire une première prédiction.

Les prévisions dans le cadre non linéaire ne sont pas aussi simples que dans le cas linéaire. Elles requièrent une gymnastique théorique qui laisse place à la créativité. Diverses méthodes ont été élaborées. Pensons à l'espérance conditionnelle $E(y_{t+h}|\Omega_t) = \int_{-\infty}^{\infty} E[y_{t+h}|y_{t+h-1}]g(y_{t+h-1}|\Omega_t)dy_{t+h-1}$. Cela a toutefois le désavantage qu'il faille évaluer des intégrales successivement, ce qui n'est pas si aisé à faire numériquement. De même, les méthodes bootstrap sont problématiques dans la mesure où elles requièrent de calculer un grand nombre de fois les coefficients, ce que nos fastidieux algorithmes de minimisation ne nous permettent pas. Évidemment, un choix s'impose et nous avons retenu une méthode. Elle est la même que celle employée par Stock et Watson (2002a) et par Shintani (2005). Elle consiste à calculer directement les régressions avec l'horizon de prédiction h. Cette approche est valide dans la mesure où on reconnaît qu'il ne s'agit là que d'une approximation et non pas du processus générateur des données.

Une seconde possibilité aurait été de considérer la méthode itérée. Cela aurait consisté à poser $h = 1$ et à évaluer notre modèle ainsi. Nous aurions ensuite pu obtenir la prévision subséquente en nous servant de la prévision que l'on vient d'obtenir et de continuer ainsi jusqu'à ce qu'on atteigne notre véritable horizon h. On rencontre une difficulté en ce qu'on doive également faire des prédictions pour les indices de diffusion. Il aurait alors fallu spécifier un modèle pour obtenir les facteurs jusqu'à $h-1$. Cela n'était toutefois pas suffisant pour écarter approche, mais on aura quand même chois de la mettre de côté, car des résultats préliminaires suggèrent que cette méthode détériore singulièrement la prédiction. Il paraît prématuré de poser un jugement définitif à savoir pourquoi cette méthode ne donnait pas les bons résultats. Il se peut que notre modèle ne soit pas assez parcimonieux et qu'il épouse le bruit de sorte que lorsqu'on l'utilise à plusieurs reprises, on obtienne des résultats insensés.

3 Résultats expérimentaux

3.1 Statistiques

Il existe plusieurs manières d'évaluer la précision de nos prévisions. Notre dévolu s'est porté sur l'erreur quadratique moyenne de prédiction pour discriminer les modèles entre eux. En utilisant la notation utilisée précédemment, celle-ci est définie par $EQMP = \frac{1}{P} \sum_{t=R}^{T-h} (y_{t+h} - \hat{y}_{t+h|t})^2$ où y_{t+h} est la valeur réalisée que l'on cherche à prédire et $\hat{y}_{t+h|t}$ est la valeur prédite par le modèle sachant l'information jusqu'à la période t. De toute évidence, l'erreur quadratique moyenne est toujours positive et notre modèle est meilleur à mesure que l'EQMP tend vers zéro. À l'aide de cette mesure, Stock et Watson (2002a) ont montré que leur modèle à facteurs est généralement supérieur à des modèles AR et VAR standards. Pour ses travaux, Shintani (2005) a utilisé l'EQMP et l'erreur absolue moyenne de prédiction définie comme $EAMP = \frac{1}{P} \sum_{t=R}^{T-h} \left| y_{t+h} - \hat{y}_{t+h|t} \right|$. On cherche alors à montrer que les méthodes non linéaires retenues sont supérieures au modèle linéaire de Stock et Watson. Le premier indicateur que l'on utilisera sera simplement le ratio entre l'erreur quadratique moyenne de prédiction du modèle linéaire et celle du modèle non linéaire. Nous aurons l'EQMP du modèle non linéaire au numérateur et celle du modèle linéaire au dénominateur. Ainsi, lorsque ce rapport sera inférieur à un, on aura une indication que le modèle non linéaire est meilleur que celui qui n'est que linéaire.

$$\frac{EQMP_{nonlin}}{EQMP_{lin}} < 1 \quad \Leftrightarrow \quad EQMP_{nonlin} < EQMP_{lin}$$

Inversement, si ce ratio est près d'un, on pourra en conclure qu'ajouter la non-linéarité n'accroît pas la précision de nos prévisions. Cette statistique a certes l'avantage d'être simple, mais elle comporte l'inconvénient de ne pas indiquer si cette différence est significative. En effet, il se peut que l'erreur quadratique moyenne de prédiction pour le modèle non linéaire soit inférieure à celle du modèle linéaire, mais que cette différence ne soit due qu'au hasard. Stock et Watson contournent cette difficulté en calculant un écart-type de cette statistique selon une méthode définie par West (1996). Ils sont alors plus en mesure de déterminer si l'écart obtenu est important en regard de cette mesure de dispersion. Pour notre part, nous utiliserons un test formel plutôt qu'un écart-type qui ne fait que nous donner qu'une vague idée. Ce test est tiré de Clark et West (2006) et se décline comme suit :

$$H_0 : \quad EQMP_{nonlin} = EQMP_{lin}$$

$$H_1 : \quad EQMP_{nonlin} < EQMP_{lin}$$

L'hypothèse alternative est unilatérale, car on cherche à prouver que le modèle non linéaire a une erreur quadratique moyenne de prédiction plus petite que celle du modèle linéaire. Ce test sert à comparer deux modèles emboîtés [10]. Dans notre cas, le modèle linéaire est emboîté dans le modèle non linéaire. Sous H_0, nous avons que le modèle non linéaire cherche à évaluer des coefficients qui sont nuls. Il se trouve ainsi à introduire du bruit dans ses prévisions. Cela a pour corollaire que le modèle plus parcimonieux aura une erreur quadratique moyenne de prédiction plus petite que le modèle plus étendu. Clark et West suggèrent donc de modifier les statistiques usuelles

$$EQMP_{lin} = \frac{1}{P} \sum_{t=R}^{T-h} (y_{t+h} - \hat{y}^{lin}_{t+h|t})^2$$

$$EQMP^*_{nonlin} = \frac{1}{P} \sum_{t=R}^{T-h} (y_{t+h} - \hat{y}^{nonlin}_{t+h|t})^2 - \frac{1}{P} \sum_{t=R}^{T-h} (\hat{y}^{lin}_{t+h|t} - \hat{y}^{nonlin}_{t+h|t})^2$$

On remarque que l'erreur quadratique moyenne de prédiction pour le cas linéaire demeure la même, mais qu'on a retranché un terme à celle du cas non linéaire. Cela réduit donc celle qu'on aurait obtenue si on avait utilisé la statistique habituelle et augmente donc les chances de rejeter H_0. Les auteurs dont il est question trouvent que la nouvelle statistique associée à ce test n'a pas une distribution normale, mais qu'elle est tout de même assez proche. En fait, ils ont calculé que les tests ont des niveaux plus petits que ceux spécifiés par l'approximation normale [11]. En conséquence de quoi, lorsqu'on en vient à rejeter H_0 dans ce test, nous sommes davantage certains que cette hypothèse est erronée que nous le suggère la statistique obtenue.

On peut alors tester en régressant par moindres carrés ordinaires la différence entre les erreurs quadratiques moyennes pour les deux modèles. Notre test consiste en un test Student unilatéral en se servant de l'approximation normale. On rapportera ensuite la valeur p obtenue à l'aide de l'approximation normale.

Par ailleurs, on notera que les réseaux de neurones sont souvent considérés comme des modèles « boîte noire » et l'on s'en sert principalement pour leurs

10. Le modèle 1 est emboîté (*nested* en anglais) dans le modèle 2 si et seulement si certains coefficients du modèle 1 sont considérés comme nuls, alors qu'ils ne le sont pas pour le modèle 2. On aura ainsi que le modèle 1 est un cas particulier du modèle 2.

11. En comparaison, avec les statistiques traditionnelles, ils obtiennent des tests qui sous-estiment grossièrement les niveaux quand on utilise l'approximation normale.

qualités prédictives. Par « boîte noire », on entend qu'il est peu commode d'analyser le comportement interne du modèle. En effet, les propriétés asymptotiques habituelles ne sont pas valides dans ce cadre-ci. Par exemple, on ne peut pas recourir à la statistique de Student pour tester les coefficients de la partie non linéaire. On se contente alors d'analyser l'erreur quadratique moyenne de prédiction. Ainsi, pour ce qui est des réseaux de neurones, notre analyse se limitera aux deux statistiques énoncées précédemment, à savoir le ratio des erreurs quadratiques moyennes de prédiction et la valeur p. Cela dit, on remarquera que le choix des réseaux neuronaux conduit à une heureuse coïncidence des inconvénients. Il s'avère qu'il n'est pas bien clair ce que représente un facteur. De même, on ne parvient pas à déterminer ce que signifient les coefficients. Cela fait en sorte qu'on ne sait pas interpréter les coefficients associés à des variables qu'on ne sait trop ce qu'elles sont. Dans un cadre strictement de prédiction, ces deux points faibles perdent de leur portée.

En ce qui a trait aux séries de Fourier, on utilisera également les deux statistiques que l'on emploiera pour les réseaux de neurones. De plus, il aurait pu être envisagé de tester les coefficients de la partie non linéaire avec un test F habituel. En effet, le modèle linéaire et celui non linéaire sont tous les deux évalués avec la méthode des moindres carrés ordinaires, ce qui autorise l'utilisation d'un test F usuel. Seulement, la sélection du meilleur modèle à l'aide du critère du BIC rend inopérant cette possibilité dans la mesure où les variables incluses de la partie linéaire ne correspondent pas nécessairement pour les deux modèles.

3.2 Résultats empiriques

Les résultats sont présentés sous forme de tableau. Pour chaque variable, on trouvera dans la case de gauche le ratio des erreurs quadratiques moyennes de prédiction entre le modèle à gauche et celui considéré par Stock et Watson (2002a). Dans la case de droite, on trouvera la valeur p associée au test énoncé précédemment. Pour le ratio, une valeur inférieure à un signifie qu'il y a une amélioration par rapport au modèle linéaire. De même, on pourra considérer qu'une valeur p inférieure à 0,05 est une amélioration vis-à-vis du modèle linéaire. On aura mis ces valeurs en gras dans le tableau.

Le premier tableau reprend quatre variables réelles étudiées par Stock et Watson pour des horizons de 6 mois, 12 mois et 24 mois. Le deuxième tableau rapporte les valeurs pour les variables de prix qu'ont étudié ces auteurs.

TABLE 1 – EQMP relative et la valeur p pour des variables réelles

Méthode	Prod. indust.		Rev. pers.		Ventes		Salariés	
RN (6 mois)	1,21	0,95	1,05	0,58	**0,99**	0,07	1,09	0,29
RN (12 mois)	1,11	0,23	1,07	0,61	1,00	**0,04**	1,09	0,29
RN (24 mois)	1,07	**0,02**	1,09	0,63	1,09	0,95	1,14	0,20
Fourier (6 mois)	**0,97**	**< 0,01**	**0,96**	**< 0,01**	1,01	0,14	**0,90**	**< 0,01**
Fourier (12 mois)	1,04	0,35	1,04	0,69	**0,99**	**0,05**	1,00	**0,01**
Fourier (24 mois)	1,04	0,95	1,02	0,61	1,04	0,45	1,06	0,31

Légende : Prod. indust. = production industrielle, Rev. pers. = revenu personnel, Ventes manuf. = ventes manufacturières et commerciales, Salariés = nombre de salariés excluant la population agricole.

TABLE 2 – EQMP relative et la valeur p pour des variables de prix sur un horizon de 12 mois

Méthode	IPC		Défl. Cons.		IPC		IPP	
RN (6 mois)	1,00	0,04	1,07	0,96	**0,96**	**< 0,01**	1,00	1,00
RN (12 mois)	1,03	0,30	1,03	0,51	1,04	0,52	1,06	0,75
RN (24 mois)	1,04	0,18	1,04	0,59	1,05	0,49	1,07	0,73
Fourier (6 mois)	1,01	0,47	1,00	0,39	1,01	0,67	**0,99**	0,15
Fourier (12 mois)	1,00	0,56	1,00	0,32	1,00	0,32	1,00	0,30
Fourier (24 mois)	**0,99**	0,09	1,00	0,25	1,00	0,28	**0,99**	**0,04**

Légende : IPC = indice des prix à la consommation, Défl. Cons. = déflateur implicite des prix à la consommation, IPC = IPC excluant la nourriture et l'énergie, IPP = indice des prix à la production.

3.3 Analyse des résultats

En regardant les ratios, nous sommes à même de constater que la plupart d'entre eux sont supérieurs à un. D'un point de vue strictement prévisionnel, nous sommes amenés à en déduire que les méthodes non linéaires détériorent les prévisions par rapport au modèle linéaire. Certes, il arrive à l'occasion que le ratio soit inférieur à un indiquant ainsi qu'il y aurait des non-linéarités, mais cela n'est pas assez répandu pour en conclure à la présence de non-linéarité. De plus, la majorité des valeurs ne sont que légèrement supérieures à un de sorte que les modèles non linéaires ne nuisent pas outrageusement à nos prédictions. Cela est bien naturel dans la mesure où ne fait qu'ajouter une partie non linéaire à des régresseurs linéaires. Dans une perspective strictement prédictive, il ne semble dès lors pas pertinent d'inclure ces méthodes non linéaires à la partie

linéaire déjà considérée par Stock et Watson. La seule exception apparente serait peut-être les séries de Fourier avec un horizon court et pour les variables réelles. Cela n'est toutefois pas suffisant pour nous amener à induire qu'il y a des non-linéarités. Du reste, il ne semble pas bien clair ce qui expliquerait cette bonne performance.

Pour ce qui est du test qui nous indique si les erreurs quadratiques moyennes entre les modèles linéaire et non linéaire sont différentes, il semble que là aussi nous manquons de preuve pour attester de la présence de non-linéarité. Relativement peu de valeurs-p sont sous $0,05$, ce qui atteste que la méthode linéaire serait probablement suffisante pour établir les prédictions pour les modèles à facteurs. Bref, même en tenant compte de l'incertitude en évaluant les coefficients supplémentaires du modèle non linéaire, nous ne sommes pas en mesure de conclure que les méthodes non linéaires nous permettent d'améliorer les prévisions.

Dès lors, à la lumière de ces résultats, il ne semble pas pertinent d'inclure ces méthodes non linéaires à la partie linéaire déjà considérée par Stock et Watson. Il se trouve que Shintani (2005) n'arrive pas non plus à trouver que les réseaux de neurones améliorent les prévisions bien qu'il ait trouvé quelques indices de non-linéarité dans les données à partir de divers tests. Il était par ailleurs arrivé à la conclusion que les indices de diffusion amélioraient les prévisions par rapport aux modèles linéaires. Comme nous utilisons les données de Stock et Watson et que cette dernière conclusion avait déjà été démontrée par ceux-ci, notre contribution se limite alors à indiquer que les méthodes non linéaires étudiées ne permettent pas d'améliorer les modèles incluant des indices de diffusion.

Il serait par la suite naturel de se demander ce qui fait en sorte que les méthodes non linéaires ne nous confèrent aucun avantage apparent par rapport aux méthodes linéaires. Plusieurs interprétations s'offrent alors à nous pour expliquer nos résultats. Tout d'abord, il se peut qu'il n'y ait tout simplement pas de non-linéarité. Il s'avérerait alors futile de vouloir considérer les modèles non linéaires. C'est l'explication la plus plausible et celle que l'on assumera. Ensuite, il se peut qu'il y ait de la non-linéarité, mais que les méthodes que nous avons utilisées, soit les réseaux de neurones et les séries de Fourier, soient mal adaptées pour l'expliquer. Également, nous avons fait mention précédemment du fait que le temps de calcul trop long nous a indûment restreints pour ce qui est de la précision du modèle incorporant des réseaux de neurones. En effet, pour évaluer les coefficients, nous avons eu recours à une approximation qui a pu parfois s'avérer inadaptée. Nous avons employé deux algorithmes avec un nombre donné d'itérations pour chacun de sorte que nous ne sommes pas en mesure d'affirmer que nous avons à tout coup trouvé le minimum. De même, nous avons peu d'assurance à savoir si nous avons trouvé des minimums globaux ou des minimums locaux. De plus, nous avons fixé tous les paramètres du modèle avec les réseaux de neurones, ce qui est très limitatif par rapport au modèle linéaire qui emploie le BIC. Toutefois, on notera que ces limitations perdent de leur portée dans la

mesure où le modèle avec les séries de Fourier n'a pas ces contraintes et qu'il ne performe guère mieux que celui avec les réseaux neuronaux. Enfin, une dernière explication qui paraît assez peu concluante est que la valeur p associée à la statistique a des niveaux plus petits que ceux prédits. Il serait alors possible que les valeurs p que l'on obtient soient significatives si nous avions un test qui avait les bons niveaux. Cela ne change toutefois pas le fait que les méthodes non linéaires n'améliorent pas concrètement les prévisions simplement à cause de l'évaluation des coefficients supplémentaires.

4 Conclusion

4.1 Sommaire

Dans ce rapport, nous avons considéré étendre le modèle considéré par Stock et Watson (2002a) à l'aide de méthodes non linéaires. Comme ceux-ci, nous avons procédé en deux étapes pour obtenir notre modèle à facteurs dynamiques. Tout d'abord, nous avons estimé les facteurs à partir de 146 séries en utilisant la méthode des composantes principales. Ensuite, nous avons inclus ces prévisions dans des modèles non linéaires pour les comparer au modèle linéaire de Stock et Watson. Nous avons utilisé les réseaux de neurones ainsi que les séries de Fourier. Dans le premier cas, nous avons dû fixer les paramètres de notre modèle, car le temps de calcul était trop long. Pour les séries de Fourier, nous avons utilisé le BIC tout comme les auteurs l'avaient fait avec leur modèle linéaire. De plus, les prévisions sont faites directement plutôt que de manière itérée.

Pour déterminer si des non-linéarités étaient présentent, nous avons calculé le ratio des erreurs quadratiques moyennes de prévision entre le modèle non linéaire et celui qui est linéaire. Nous avons aussi modifié l'erreur quadratique moyenne de prévision du modèle non linéaire, ce qui nous permet d'effectuer un test qui détermine si l'EQMP du modèle non linéaire est inférieure à celle du modèle linéaire.

À la lumière de nos résultats, nous ne sommes pas en mesure d'affirmer qu'il serait pertinent de modifier le modèle de Stock et Watson (2002a) afin d'inclure une partie non linéaire. Si l'on ne considère que l'aspect prédictif, il semble que la partie non linéaire améliore rarement les prévisions, que plus souvent qu'autrement, elle ne change rien et qu'il lui arrive parfois de détériorer les prévisions. Même lorsque l'on considère l'erreur associée au fait de calculer des coefficients supplémentaires, nous ne pouvons pas dire que les modèles non linéaires améliorent nos prévisions.

4.2 Améliorations possibles

Cette recherche est sujette à quelques améliorations qu'il importe d'énoncer pour guider un travail subséquent si tant est que cette avenue s'avère suffisamment prometteuse. Tout d'abord, il pourrait être avantageux de permettre toutes les complexités associées au modèle avec les réseaux de neurones. Cela prendrait nécessairement plus de temps, mais il se peut que des gains en termes de prévision se trouvent de ce côté. Ensuite, il serait peut-être souhaitable d'explorer d'autres méthodes non linéaires afin de voir si l'apparente absence de non-linéarité ne serait pas due à un mauvais choix de modèles non linéaires. On pourrait par exemple considérer d'intégrer les facteurs dans des vaguelettes [12]. De même, il pourrait être judicieux de considérer des tests de non-linéarité. Si ceux-ci nous indiquent que les données ne sont pas linéaires, il pourrait alors s'avérer pertinent de considérer d'autres méthodes non linéaires. Shintani (2005) a procédé à certains tests et il a effectivement trouvé des indices de non-linéarités. Il a néanmoins trouvé peu de gains des réseaux de neurones par rapport au modèle linéaire. Enfin, des études empiriques avec d'autres données s'avéreraient utiles puisque nos résultats ne sont valides que pour les États-Unis. Shintani a toutefois pris des données pour le Japon et est arrivé à la même conclusion que nous.

5 Annexe

5.1 Algorithme BFGS

L'algorithme BFGS s'apparente à l'algorithme de minimisation de Newton. On part de l'approximation de Taylor d'ordre 2 de $Q(\theta)$ qui est $Q(\theta) = Q(\theta_n) + (\theta - \theta_n) \cdot \nabla Q(\theta_n) + \frac{1}{2}(\theta - \theta_n) \cdot A \cdot (\theta - \theta_n)$ où A est la matrice hessienne. Lorsqu'on dérive cette fonction par rapport à θ, on obtient $\nabla Q(\theta) = \nabla Q(\theta_n) + A \cdot (\theta - \theta_n)$. Avec la méthode de Newton, on pose $\nabla Q(\theta) = 0$ puisqu'on recherche un minimum, ce qui nous donne le point pour la prochaine itération, c'est-à-dire $\theta = \theta_n - A^{-1} \cdot \nabla Q(\theta_n)$. L'algorithme BFGS utilise, à la place de la matrice hessienne, une suite d'approximations de l'inverse de celle-ci $\{H_n\}_{n=1}^{\infty}$ qui est telle que $\lim_{n \to \infty} H_n = A^{-1}$. Cela lui confère une vitesse de convergence de l'ordre n^2. À partir d'une première approximation, on calcule les points $\{\theta_n\}_{n=1}^{\infty}$ successivement jusqu'à ce qu'on ait atteint la précision voulue. À chaque itération, on trouve la direction $p_n = H_n \nabla Q(\theta_n)$ dans laquelle on doit chercher le prochain point. On cherche le minimum de $Q(\theta)$ dans la direction de p_n à partir de θ_n, ce qui nous donne θ_{n+1}. Ainsi, la méthode BFGS s'avère parfois supérieure à celle de Newton étant donné que cette manière de procéder nous conduit nécessairement à faire décroître la

12. Traduction de *wavelets*.

fonction, alors que ce n'est pas nécessairement le cas avec la matrice hessienne. Ensuite, on calcule l'approximation suivante de l'inverse de la matrice hessienne. La suite de matrices est construite ainsi :

$$H_{n+1} = H_n + ((\nabla Q_{n+1} - \nabla Q_n) \cdot H_n \cdot (\nabla Q_{n+1} - \nabla Q_n))\, u \times u \\ + \frac{(\theta_{n+1}-\theta_n)\times(\theta_{n+1}-\theta_n)}{(\theta_{n+1}-\theta_n)\cdot(\nabla Q_{n+1}-\nabla Q_n)} - \frac{(H_n\cdot(\nabla Q_{n+1}-\nabla Q_n))\times(H_n\cdot(\nabla Q_{n+1}-\nabla Q_n))}{(\nabla Q_{n+1}-\nabla Q_n)\cdot H_n\cdot(\nabla Q_{n+1}-\nabla Q_n)}$$

avec

$$u = \frac{\theta_{n+1} - \theta_n}{(\theta_{n+1} - \theta_n) \cdot (\nabla Q_{n+1} - \nabla Q_n)} - \frac{H_n \cdot (\nabla Q_{n+1} - \nabla Q_n)}{(\nabla Q_{n+1} - \nabla Q_n) \cdot H_n \cdot (\nabla Q_{n+1} - \nabla Q_n)}$$

où $\nabla Q_n = \nabla Q(\theta_n)$ et $\times$ est le produit vectoriel. On pose généralement $H_1 = I$. Enfin, on va se contenter d'énoncer les formules nécessaires pour implanter l'algorithme, car il s'avérerait fastidieux et superfétatoire d'expliquer en détail la justification de cette approche. Un lecteur intéressé pourra consulter Kelly (1999).

6 Bibliographie

Alonso, Andre'es M., Daniel Peña et Juan Romo (2002), « Forecasting time series with sieve bootstrap », *Journal of Statistical Planning and Inference*, vol. 100, no. 1, p. 1-11.

Bai, Jushuan et Serena Ng (2002a), « Determining the number of factors in approximate factor models », *Econometrica*, vol. 70, no. 1, p. 191-221.

Bai, Jushuan et Serena Ng (2008), « Large Dimensional Factor Analysis », *Foundations and Trends in Econometrics*, vol. 3, no. 2, p. 89-163.

Bierens, Herman, Ivan Castelar, Roberto Tatiwa Ferreira (2005), « Forecasting Quartely Brazilian GDP Growth Rate With Linear and Non-linear Diffusion Index Models », *Revista EconomiA*, vol. 6, no. 3. p. 261-292.

Bloomfield, Peter (2000), *Fourier Analysis of Time Series : An Introduction*, Deuxième édition, Wiley-Interscience, New York, 288 p.

Burns, Arthur F. et Wesley C. Mitchell (1947), *Measuring Business Cycles*, New York, National Bureau of Economic Research, 590 p.

Chen, Xiaohong, Jeffrey Racine et Norman R. Swanson (2001), « Semiparametric ARX Neral-network Models with an Application to Forecasting Inflation », *IEEE Transactions on Neural Networks*, vol. 12, no. 4, p. 674-683.

Clark, Todd E. et Kenneth D. West (2006), « Approximately normal tests for equal predictive accuracy in nested models », *Journal of Econometrics*, vol. 138, no. 1, p. 291-311.

Engle, Robert et Mark Watson (1981), , « A One-Factor Multivariate Time Series Model of Metropolitan Wage Rates », *Journal of the American Statistical Association*, vol. 76, no. 376, p. 774-781.

Flannery, Brian P., William H. Press, Saul A. Teukolski et William T. Vetterling (1992). *Numerical Recipes in C*, Deuxième édition, Cambridge University Press, Cambridge, 994 p.

Forni, Mario et al. (1998), , « Let's Get Real : A Factor Analytical Approach to Disaggregated Business Cycle Dynamics », *Review of Economic Studies*, vol. 65, no. 3, p. 453-473.

Forni, Mario et al. (2000), , « The Generalized Dynamic-Facotr Model : Identification and Estimation », *The Review of Economics and Statistics*, vol. 82, no. 4, p. 540-554.

Franses, Philip Hans et Dick van Dijk (2000), *Non-Linear Time Series Models in Empirical Finance*, Cambridge, Cambridge University Press, 280 p.

Geweke, John F. et Kenneth J. Singleton (1981), « Measuring the Princing Error of the Arbitrage Pricing Theory », *Review* of Financial Studies, vol. 9, no. 2, p. 557-587.

Geweke, John F. et Guofu Zhou (1996), « Maximum Likelihood "Conrimatory" Factor Analysis of Economic Time Series », *International Economic Review*, vol. 22, no. 1, p. 37-54.

Gallant, A. Ronald (1981), « On the Bias in Flexible Fuctinoal Forms and an Essentially Unbiased Form : The Fourier Flexible Form », *Journal of Econometrics*, vol. 15, no. 2, p. 211-245.

Giorano, Francesco, Michele La Rocca et Cira Perna (2006), « Forecasting nonlinear time series with neural network sieve bootstrap », *Computational Statistics & Data Analysis*, vol. 51, no. 8, p. 3871-3884.

Hong, Yongmiao et Tae-Hwy Lee (2003), « Inference on Predictability of Foreign Exchange Rates via Generalized Spectrum and Nonlinear Time Series Models », *Review of Economics and Statistics*, vol. 85, no. 4, p. 1048-1062.

Kelley, C. T. (1999), *Iterative Methods for Optimization*, Philadelphia, Society for Industrial and Applied Mathematics, 180 p.

KIM, Chang-Jin et Charles R. NELSON (1998), « Business Cycle Turning Points : A New Coincident Index and Tests of Duration Dependence Based on a Dynamic Factor Model with Regime-Switching », *Review of Economics and Statistics*, vol. 80, no. 2, p. 188-201.

LINTON, Oliver et Benoit PERRON (2003), « The Shape of the Risk Premium : Evidence From a Semiparametric Generalized Autoregressive Conditional Heteroscedasticity Model », *Journal of Business & Economic Statistics*, vol. 21, no. 3, p. 354-367.

MARCELLINO, Massimiliano, James H. STOCK et Mark W. WATSON (2003), « Macroeconomic forecasting in the Euro area : Country specific versus area-wide information », European Economic Revi*ew*, vol. 47, no. 1, p. 1-18.

NELDER, J. A. et R. MEAD (1965), « A simplex method for fuction minimization », *Computer Journal*, vol. 7, no. 4, p. 308-313.

SARGENT, Thomas J. (1989), « Two Models of Measurements and the Investment Accelerator », *Journal of Political Economy*, vol. 97, no. 2, p. 251-287.

SARGENT, Thomas J. et SIMS Christopher A. (1977), « Business Cycle Modeling without Pretending to Have Too Much A-Priori Economic Theory », Federal Reserve Bank of Minneapolis, Working Papers, no. 55.

SHINTANI, Mototsugu (2005), « Nonlinear Forecasting Analysis Using Diffusion Indexes : An Application to Japan », *Journal of Money, Credit, and Banking*, vol. 37, no. 3, p. 517-538.

STOCK, James H. et Mark W. WATSON (1989), « New Indexes of Coincident and Leading Economic Indicators », *NBER Macroeconomics Annual 1989*, vol. 4, p. 351-409.

STOCK, James H. et Mark W. WATSON (1991), « A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems », *Econometrica*, vol. 61, no. 4, p. 819-840.

STOCK, James H. et Mark W. WATSON (1998), « Diffusion Indexes », *National Bureau of Economic Research*, Technical woking paper 6702.

STOCK, James H. et Mark W. WATSON (2002a), « Macroecnomic Forecasting Using Diffusion Indexes », *Journal of Business & Economic Statistics*, vol. 20, no. 2, p. 147-162.

STOCK, James H. et Mark W. WATSON (2002b), « Forecasting Using Principal Components From a Large Number of Predictors », *Journal of Business & Economic Statistics*, vol. 97, no. 460, p. 1167-1179.

Stock, James H. et Mark W. Watson (2006), « Forecasting With Many Predictors », dans Graham Elliott, Clive W.J. Granger et Allan Timmermann (dir.), *Handbook of Economic Forecasting*, vol. 1, chapitre 10, Amsterdam, North-Holland, p. 515-554.

Swanson, Norman R. et Halbert White (1997), « A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks », *Review of Economics and Statistics*, vol. 79, no. 4, p. 540-550.

West, Kenneth D. (1996), « Asymptotic Inference about Predictive Ability », *Econometrica*, vol. 64, no. 5, p. 1067-1084.

West, Kenneth D. (2006), « Forecast Evaluation », dans Graham Elliott et Clive W.J. Granger et Allan Timmermann (dir.), *Handbook of Economic Forecasting*, vol. 1, Amsterdam, North-Holland, p. 99-134.