Introduction to multilayer perceptron and hybrid hidden Markov, multilayer perceptron, models.

27 juin 2002

1 Introduction.

Une série temporelle est une suite de mesures équi-espacées dans le temps. C'est par exemple les cours du CAC 40, jour après jour, le PIB d'un pays année après année. Il est évident que si l'on pouvait deviner le comportement de ces séries, on pourrait en tirer un grand avantage. Pour cela, les statisticiens ont développé des outils pour modéliser ce comportement et essayer de prévoir au mieux les valeurs futures du processus observé. Nous allons étudier dans ce document l'apport des réseaux de neurones et plus particulièrement des perceptrons multicouches (MLP) pour les séries temporelles. Dans la première partie, nous aborderons essentiellement la sélection d'architecture des MLP, qui permettra d'éviter la surparamétrisation du modèle, qui conduit au phénomène de sur-apprentissage. Dans la seconde partie, nous traiterons le cas de séries stationnaires par morceaux qui exigent d'utiliser plusieurs modèles de régression simultanément et de choisir, de façon probabiliste, à chaque instant, celui qui fait la prédiction la plus pertinente. Finalement, nous appliquerons ces méthodes à la modélisation d'une série de pollution en niveau d'ozone à Paris.

2 Modèle auto-régressifs

On s'intéresse, dans ce document, à la modélisation paramétrique des séries temporelles. Plus particulièrement, nous étudions les modèles utilisant des perceptrons multicouches (MLP) comme fonction de régression.

On considère le modèle de séries temporelles suivant :

$$Y_{t+1} = F_{W_0}(Y_t) + \varepsilon_{t+1}$$

où

- $Y_t \in \mathbb{R}$ est l'observation au temps "t" de la série temporelle.
- ε_t est un bruit i.i.d. d'espérance 0, de variance constante σ^2, par exemple une variable $\mathcal{N}(0, \sigma^2)$, indépendante du passé de la série.
- F_{W_0} est une fonction représentée par un MLP ayant pour paramètres (poids) le vecteur $W_0 \in \mathbb{R}^D$.

Pour simplifier l'écriture, on ne considère que des modèle autorégressifs d'ordre 1, cependant la généralisation à un ordre supérieur est très facile. Le phénomène observé (Y_t) est donc la combinaison d'un fonction déterministe du passé du processus et d'un aléa. Si on connaît la fonction déterministe F_{W_0} sous-jacente, on peut faire les meilleures prévisions possibles. Comme cette fonction est entièrement déterminée par son vecteur paramètre, le travail du statisticien consiste donc à estimer le paramètre W_0 à l'aide d'un nombre fini d'observations $(y_0, \cdots, y_T)$. Pour cela on minimise en W une fonctionnelle comme

$$S_T(W) = \frac{1}{T}\sum_{t=1}^{T}\left(Y_t - F_W\left(Y_{t-1}\right)\right)^2$$

la moyenne des carrés résiduels et on note

$$\hat{W}_T = \arg\min_W S_T(W)$$

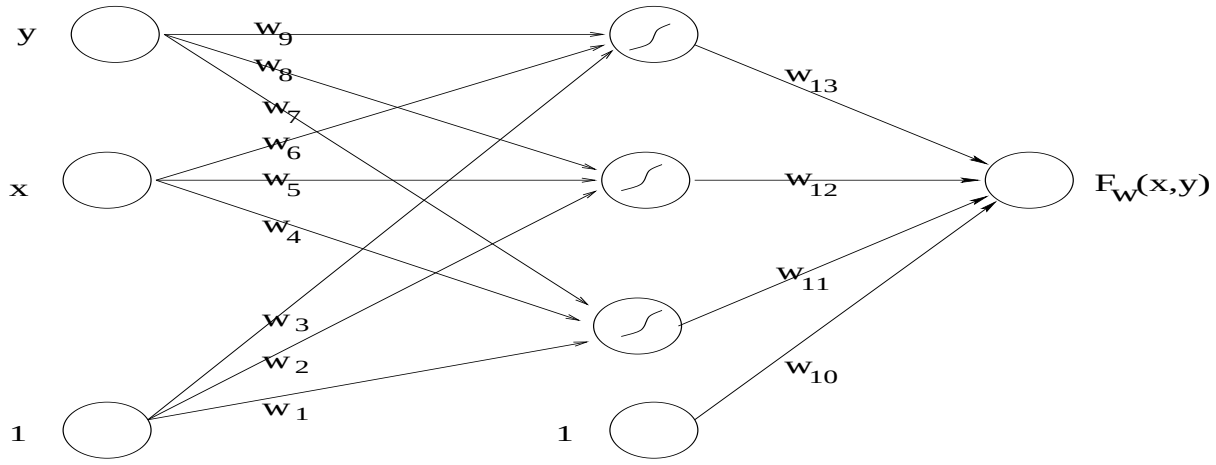
l'estimateur des moindres carrés de W_0.

2.1 Résultats théoriques (MLP à une couche cachée)

2.1.1 Le modèle MLP

On peut représenter une fonction par un MLP de la façon suivante :

FIG. 1 – Perceptron multicouche



Le MLP ici représenté est une fonction de $\mathbb{R}^2 \to \mathbb{R}$ qui à (x, y) associe $F_W(x, y)$ avec

$$F_W(x,y) = w_{10} + w_{11}\phi\left(w_1 + x \times w_4 + y \times w_7\right) + w_{12}\phi\left(w_2 + x \times w_5 + y \times w_8\right)$$
$$+w_{13}\phi\left(w_3 + x \times w_6 + y \times w_9\right)$$

La fonction d'activation ϕ de la couche cachée est généralement une fonction sigmoïde, que l'on pourra prendre, sans perte de généralité, égale à la tangente hyperbolique. Le vecteur paramètre est donc ici $W = (w_1, \cdots, w_{13})$.

Dans toute la suite, on supposera que notre modèle est identifiable, c'est-à-dire que pour une fonction représentable par un MLP donné, il n'y a qu'un seul vecteur paramètre qui représente cette fonction. Pour obtenir cette propriété, on doit restreindre l'ensemble des paramètres convenablement (cf Sussmann [18]).

2.1.2 Propriétés statistiques

On s'intéresse au comportement de l'estimateur $\hat{W}_T$ lorsque T tend vers l'infini. Il est intéressant d'avoir deux propriétés fondamentales :

- La consistance, c'est-à-dire que $\hat{W}_T \overset{T\to\infty}{\longrightarrow} W_0$
- La normalité asymptotique, qui assure que la convergence précédente se fait à la vitesse $\sqrt{T}$, et permet d'obtenir la loi limite de $\hat{W}_T$. On trouvera par exemple dans Yao [19] la démonstration du théorème suivant :

Théorème 1 *Consistance et normalité asymptotique de l'estimateur $\hat{W}_T$ avec $\phi(x) = \tanh(x)$, supposons que :*

1. *$(\varepsilon_t)_{t\in\mathbb{N}^*}$ est une suite i.i.d. telle que $E\varepsilon_t^6 < \infty$,*
2. *W_0 appartient à l'intérieur d'un sous-ensemble compact de l'espace euclidien $\mathbb{R}^D$.*
3. *Si μ_0 est la mesure stationnaire de (Y_t), la matrice de dimension $m \times m$*

$$\Sigma_0 = \int_{\mathbb{R}} \left[\frac{\partial}{\partial w_i} F_{W_0}(y) \frac{\partial}{\partial w_j} F_W(y) \right]_{1\le i,j\le m} \mu_0(dy)$$

est définie positive.

Alors :

- *L'estimateur $\hat{W}_T$ converge presque sûrement vers W_0 quand T tend vers $+\infty$.*
- *Le terme $\sqrt{T}\left[\hat{W}_T - W_0\right]$ converge en loi vers la distribution gaussienne multidimensionnelle $\mathcal{N}\left(0, \Sigma_0^{-1}\right)$.*

2.1.3 Identification du modèle

Une des principales difficultés rencontrée lors de l'utilisation de fonctions de plus en plus complexes pour l'estimation statistique des processus est le phénomène de surdétermination des modèles (overfitting). En effet si on utilise un modèle trop complexe sur trop peu de données, on aboutit à la modélisation du bruit qui a engendré les données sur lesquelles on estime le modèle. On introduit ainsi un biais qui compromet fortement la capacité de prédiction du modèle sur de nouvelles données, non encore observées, du même processus. Un principe statistique efficace pour lutter contre le biais introduit par la complexification des modèles, est l'utilisation d'un terme de pénalisation qui est une fonction du nombre de paramètres (cf Akaike [1]).

Supposons qu'il existe une borne supérieure M pour toutes les dimensions possibles du modèle. Soit $(F_W)_{W\in\mathbb{R}^M}$ une famille de modèles dominants de dimension M, c'est-à-dire tel que le vrai paramètre W_0, de dimension B, puisse s'exprimer comme un vecteur de cette famille avec $M - B$ composantes nulles. Notons $\hat{W}_T^L$ un estimateur des moindres carrés de dimension L . Le

principe de parcimonie consiste alors à choisir l'estimateur qui minimise la nouvelle fonction de coût pénalisée :

$$CP\left(W^L\right) = \frac{S_T\left(W^L\right)}{T} + \frac{c\left(T\right)}{T} \times L$$

ou bien

$$CP^*\left(W^L\right) = \frac{\ln\left(S_T\left(W^L\right)\right)}{T} + \frac{c\left(T\right)}{T} \times L$$

où $c\left(T\right)$ est le taux de pénalisation. Si $c\left(T\right) = 2$ le contraste pénalisé CP^* est alors égal au critère AIC d'Akaike, si $c\left(T\right) = 2\ln\left(T\right)$, CP^* est égal au critère BIC de Schwarz [17]. A partir de ces définitions, on montre le résultat dont on peut trouver des preuves dans Yao [19] ou bien dans Rynkiewicz [15] :

Théorème 2 *On suppose que les conditions du théorème 1 sont vérifiées. On suppose aussi que le taux de pénalisation $c\left(T\right)$ est tel que*

$$\lim_T \frac{c\left(T\right)}{T} = 0, \quad et \quad \liminf_T \frac{c\left(T\right)}{2\ln\ln T} > \sigma^2\frac{\Lambda}{\lambda}$$

où Λ $(resp.\,\lambda)$ est la plus grande (resp. la plus petite) valeur propre de la matrice Σ_0. Alors le couple $\left(L, W_T^L\right)$ converge presque sûrement vers la vraie valeur et la vraie dimension $\left(L_0, W_0^{L_0}\right)$ du paramètre quand T tend vers ∞.

A partir de ce théorème, on peut donc proposer une méthodologie d'identification presque sûre pour déterminer le vrai modèle.

2.2 Recherche pratique du vrai modèle

2.2.1 Recherche d'un modèle dominant

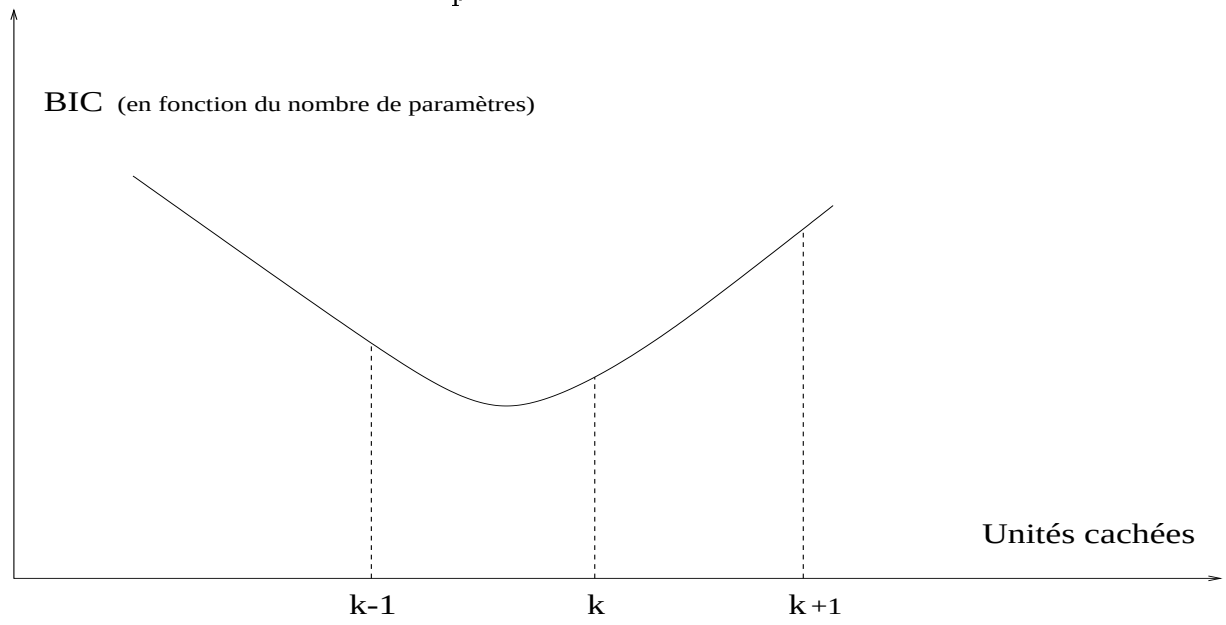
En utilisant le résultat de la section précédente, on peut donc proposer la méthode suivante de détermination du vrai modèle. Pour initialiser l'architecture, nous commençons par prendre toutes les entrées pertinentes (comme on les obtiendrait à partir d'un modèle linéaire AR) et une seule unité cachée. Ensuite, on ajoute progressivement des unités dans la couche cachée, en calculant à chaque étape le critère BIC. On continue ce processus tant que la valeur du BIC décroît. Quand l'ajout d'une unité cachée fait croître de nouveau le BIC, on arrête la recherche de modèle et on prend ce dernier MLP comme modèle dominant. On peut schématiser cette recherche par la figure suivante, qui conduit à un modèle dominant avec $k+1$ unités cachées.

2.2.2 Détermination du vrai modèle

Une fois ce modèle dominant obtenu, on obtient le vrai modèle par élagages successifs des poids.

Rappelons que $W^M = (w_1, \cdots, w_M)$ est le vecteur paramètre associé au modèle dominant. En principe, pour estimer le vrai modèle, nous devrions explorer exhaustivement la famille finie de tous les sous-modèles. Cependant ce nombre est exponentiellement grand, c'est pourquoi, comme en régression linéaire, nous proposons une méthode statistique pas à pas : Statistical Stepwise

FIG. 2 – BIC pour la recherche de modèle dominant



Method (SSM) pour guider la recherche. Une telle stratégie est basée sur le normalité asymptotique de l'estimateur $\hat{W}_T$ (cf Cottrell et al. [9]), qui utilise les statistiques de Student comme une aide à l'exploration des sous-familles du modèle dominant. Pour décider les suppressions ou non des poids w_l, on compare les valeurs des BIC du modèle F et du modèle F sans le poids w_l : F_l. Comme F_l est un sous-modèle de F il suffit que critère BIC diminue pour qu'on se rapproche un peu plus du vrai modèle, qui minimise le critère BIC. On obtient donc une suite de MLP avec de moins en moins de paramètres à laquelle correspond une trajectoire décroissante du BIC. Le critère d'arrêt de l'élagage est simple puisqu'on refuse la suppression d'un poids si celle-ci fait remonter le BIC. On garde le dernier MLP conservant ce poids.

En résumé, la procédure de recherche du vrai modèle est comme suit :

1. Déterminer F_{max} un modèle dominant
2. Grâce aux statistiques de Student, déterminer le poids l candidat à l'élimination
3. Accepter l'élimination de ce poids si et seulement si le critère BIC diminue, sinon garder pour modèle final le MLP précédant cette élagage.

Cette procédure a été testée sur de nombreux exemples et donne de très bons résultats dés que le nombre de données est suffisamment grand, généralement plus de 500 observations.

3 Modèles Hybrides

3.1 Introduction

La modélisation des séries temporelles à l'aide de réseaux de neurones permet de tenir compte d'éventuelles non-linéarités du modèle. Cependant elle repose sur l'hypothèse contraignante de la stationnarité du modèle. Une généralisation simple possible est de tenir compte des séries

stationnaires par morceaux. Hamilton [12] par exemple, a étudié de tels modèles afin de modéliser des séries temporelles sujettes à des changements discrets de régime pour analyser la série GNP (gross national product) aux Etats-Unis. On peut ainsi utiliser ces modèles par exemple pour des séries ayant un certain régime pour les périodes de croissance économique, un autre pour les périodes de récession.

Bien que ce modèle soit plus général que le précédent, on a encore besoin de nombreuses hypothèses contraignantes. D'abord le nombre de régimes possibles doit être fini. Ensuite, bien qu'il puisse y avoir des changements de régime, on suppose que ces changements interviennent de façon stationnaire, ce qui permet de récupérer, au final, une loi des grands nombres et de pouvoir ainsi faire des statistiques.

3.2 Le modèle

La théorie des chaînes de Markov cachées et leurs premières applications en reconnaissance de la parole datent de plus de 30 ans. La théorie de base a été publiée dans une suite d'articles de Baum et al. ([4, 5, 3, 2]) à la fin des années 60. Les chaînes de Markov cachées ont été utilisées par la suite dans de nombreux domaines comme la génétique, la biologie, l'économie etc...

3.2.1 Chaînes de Markov dans un espace discret

On considère $(X_t)_{t\in\mathbb{Z}}$, une chaîne de Markov homogène à valeurs dans un espace d'état fini $\mathbb{E} = \{e_1, \cdots, e_N\}$, $N \in \mathbb{N}^*$. Sans perte de généralité, on identifie l'espace d'état $\mathbb{E}$ avec le simplexe de $\mathbb{R}^N$, où e_i est un vecteur unité de $\mathbb{R}^N$ avec 1 sur la i-ème composante et 0 partout ailleurs. La chaîne X_t est caractérisée par sa matrice de transition $A = (a_{ij})_{1\leq i,j\leq N}$ qui est telle que :

$$P\left(X_{t+1} = e_i \,|X_t = e_j\right) = a_{ij}$$

Si, de plus, on définit : $V_{t+1} := X_{t+1} - AX_t$, on obtient l'écriture suivante de ce modèle :

$$X_{t+1} = AX_t + V_{t+1}.$$

3.2.2 Equations du modèle

On suppose que la série temporelle observée (Y_t) vérifie les équations suivantes :

$$\begin{cases} X_{t+1} = AX_t + V_{t+1} \\ Y_{t+1} = F_{X_{t+1}}(Y_t) + \varepsilon_{t+1}\left(X_{t+1}\right) \end{cases}$$

où $\{F_{e_1}, \cdots, F_{e_N}\}$ sont des fonctions de $\mathbb{R}^P \longrightarrow \mathbb{R}$ qui seront représentées dans notre cas par des MLP. Pour tout $e_i \in \mathbb{E}$, $(\varepsilon_t\left(e_i\right))$ est une suite de variables aléatoires indépendantes et identiquement distribuées. Ce modèle permet d'utiliser plusieurs MLP (on parle de mélange d'experts) et d'utiliser la chaîne de Markov (X_t) pour spécifier à un temps "t" quel est le MLP qui fait la prévision la plus pertinente. On remarquera que l'on observe seulement la série (Y_t), il faudra donc trouver un moyen de retrouver la suite des états de la chaîne (X_t), grâce au comportement de (Y_t).

Pour ajuster un tel modèle aux observations, on estime les paramètres (les poids des MLP F_{e_i} , la variance des bruits $\varepsilon\left(e_i\right)$, et la matrice de transition A) grâce à la méthode du maximum de vraisemblance. On pourra trouver l'étude des propriétés théoriques de cet estimateur dans Ryden et Krishnamurthy [13] et Douc, Moulines, Ryden [10].

3.3 Maximum de vraisemblance pour les modèles hybrides

Nous considérons dans la suite que la densité des bruits $(\varepsilon(e_i))_{1\leq i\leq N}$ est gaussienne. Nous commençons par étudier les paramètres libres considérés :

- La matrice de transition A, cette matrice est stochastique, c'est-à-dire que la somme d'une colonne quelconque de A est 1. Il n'y a donc que $(N-1)\times N$ paramètres libres.
- Les variances $(\sigma_{e_i})_{1\leq i\leq N}$, qui sont supposée strictement positives.
- Les paramètres des fonctions de régression $(F_{e_i})_{1\leq i\leq N}$, puisque nous utilisons des MLP, les paramètres seront clairement les vecteurs poids $(W_{e_i})_{1\leq i\leq N}$ des MLP.

Le vecteur paramètre θ sera donc :

$$\theta = \left(W_{e_1},\cdots,W_{e_N},\cdots,a_{11},\cdots,a_{(N-1)N},\sigma^2_{e_1},\cdots,\sigma^2_{e_N}\right)$$

3.3.1 Calcul de la log-vraisemblance et de sa dérivée

Nous supposerons dans la suite que la première observation y_0 ainsi que la probabilité initiale de l'état X_1 sont connues, le conditionnement des expressions par rapport à ces conditions initiales sera toujours implicite.

Une première écriture La vraisemblance du modèle pour une suite d'observations de la série $y := (y_0,\cdots,y_T)$ pour un chemin supposé réalisé $x := (x_1,\cdots,x_T)$ est donc :

$$L_\theta(y,x) = \prod_{t=1}^{T}\prod_{i=1}^{N}\left[\Phi_{e_i}(y_t - F_{e_i}(y_{t-1}))\right]^{\mathbf{1}_{\{e_i\}}(x_t)} \times \prod_{t=1}^{T}\prod_{i,j=1}^{N} a_{ij}^{\mathbf{1}_{\{e_j,e_i\}}(x_t,x_{t+1})} \times \pi_0(x_1)$$

où Φ_{e_i} est la densité la la loi normale $\mathcal{N}(0,\sigma_{e_i})$, $\mathbf{1}_G$ la fonction indicatrice de l'ensemble G et π_0 le probabilité de l'état initial x_1. Pour obtenir la vraisemblance globale des observations, on pourrait sommer ces vraisemblances sur tous les chemins possibles de la chaîne de Markov cachée. On aurait alors

$$L_\theta(y) = \sum_x L_\theta(y,x)$$

Il est bien connu que la complexité de cette somme est exponentielle, ce qui rend le calcul difficile dés que le nombre d'observations est supérieur à plusieurs centaines. Il serait aussi possible de calculer le maximum de vraisemblance grâce à l'algorithme E.M. en utilisant l'algorithme forward-backward de Baum et Welch. Cependant, nous préférons utiliser ici une technique d'optimisation différentielle qui est généralement plus rapide que l'algorithme E.M. lorsque les fonctions de régression utilisées sont des perceptrons multicouches.

La log-vraisemblance Une façon plus élégante d'écrire la log-vraisemblance est d'utiliser le filtre prédictif $P(X_t = e_i\,|y_{t-1},\cdots y_0) := p_t(i)$ puisque la vraisemblance s'écrira

$$L_\theta(y_1,\cdots,y_T) = \prod_{t=1}^{T} L_\theta(y_t\,|y_{t-1},\cdots,y_0) = L_\theta(y_t\,|y_{t-1}\cdots,y_0) \times \prod_{t=1}^{T-1} L_\theta(y_t\,|y_{t-1},\cdots,y_0)$$

$$= \sum_{i=1}^{N} L(y_T\,|X_T = e_i, y_{T-1},\cdots,y_0)P(X_T = e_i\,|y_{T-1},\cdots,y_0) \times \prod_{t=1}^{T-1} L(y_t\,|y_{-p+1},\cdots,y_{t-1}).$$

On note
- p_t le vecteur dont la i-ème composante est : $p_t(i) = P(X_t = e_i \,|y_{t-1}, \cdots, y_0)$
- b_t le vecteur dont la i-ème composante est : $b_t(i) = L(y_t \,|X_t = e_i, y_{t-1}, \cdots, y_0)$, c'est-à-dire la densité conditionnelle de y_t sachant $X_t = e_i$ et $(y_{t-1}, \cdots, y_0)$.
- u^T le vecteur u transposé.

On aura :

$$L(y_1, \cdots, y_T) = b_T^T p_T \times \prod_{t=1}^{T-1} L(y_t \,|y_{t-1}, \cdots, y_0) = \prod_{t=1}^{T} b_t^T p_t.$$

On en déduit une forme pratique de la log-vraisemblance :

$$\ln(L(y_1, \cdots, y_n)) = \sum_{t=1}^{n} \ln(b_t^T p_t). \tag{1}$$

Il suffit donc de calculer p_t pour $t = 1, \cdots, n$, pour pouvoir calculer la log-vraisemblance, car :

$$b_t(i) = L(y_t \,|X_t = e_i, y_{t-1}, \cdots, y_0) := \Phi_{e_i}(y_t - F_{e_i}(y_{t-1}))$$

Calcul de p_t En notant $B_t = diag\,(b_t)$, la matrice diagonale ayant pour diagonale le vecteur b_t, on vérifie facilement (cf Rynkiewicz [16, 15]) que le filtre prédictif p_t vérifie la récurrence :

$$p_{t+1} = \frac{AB_t p_t}{b_t^T p_t}. \tag{2}$$

On supposera que p_1 suit la distribution uniforme sur $\{1, \cdots, N\}$ et on pourra ainsi calculer p_t, $t = 1, \cdots, T$ par récurrence. Le choix de la valeur initiale de p_t a relativement peu d'importance grâce à la propriété d'oubli exponentiel de la distribution initiale (cf Legland et Mevel [14]).

3.4 Dérivée de la log-vraisemblance

On rappelle que l'on a

$$\ln(L(y_1, \cdots, y_T)) = \sum_{t=1}^{T} \ln(b_t^T p_t)$$

donc, si on note θ_j le j-ème paramètre du modèle, on a

$$\frac{\partial \ln(L(y_1, \cdots, y_n))}{\partial \theta_j} = \sum_{t=1}^{T} \frac{\frac{\partial b_t^T p_t}{\partial \theta_j}}{b_t^T p_t}.$$

Il suffit donc de calculer $\frac{\partial b_t^T p_t}{\partial \theta_j}$ pour pouvoir calculer la dérivée de la log-vraisemblance, en remarquant que :

$$\frac{\partial b_t^T p_t}{\partial \theta_j} = \frac{\partial b_t}{\partial \theta_j}^T p_t + b_t^T \frac{\partial p_t}{\partial \theta_j}. \tag{3}$$

On pourra trouver les détails du calcul de cette dérivée dans Rynkiewicz [15]. On va juste expliciter ici le calcul de la dérivée du filtre :

Calcul de $\frac{\partial p_t}{\partial \theta_j}$ suivant θ_j Comme on a la récurrence

$$p_{t+1} = \frac{AB_t p_t}{b_t^T p_t}$$

En dérivant cette expression par rapport au paramètre θ_j, on aura :

$$\frac{\partial p_{t+1}}{\partial \theta_j} = \frac{\partial AB_t p_t}{\partial \theta_j} \times \frac{1}{b_t^T p_t} + AB_t p_t \times \frac{\partial b_t^T p_t}{\partial \theta_j} \times \left(-\frac{1}{(b_t^T p_t)^2} \right)$$

soit

$$\frac{\partial p_{t+1}}{\partial \theta_j} = \left(\frac{\partial AB_t}{\partial \theta_j} p_t + AB_t \frac{\partial p_t}{\partial \theta_j} \right) \times \frac{1}{b_t^T p_t} + AB_t p_t \times \left(\frac{\partial b_t^T}{\partial \theta_j} p_t + b_t^T \frac{\partial p_t}{\partial \theta_j} \right) \times \left(-\frac{1}{(b_t^T p_t)^2} \right).$$

On a alors :

$$\frac{\partial p_{t+1}}{\partial \theta_j} = \frac{AB_t}{b_t^T p_t} \left[I - \frac{p_t b_t^T}{b_t^T p_t} \right] \frac{\partial p_t}{\partial \theta_j} + \left(\frac{\partial AB_t}{\partial \theta_j} \right) \frac{p_t}{b_t^T p_t} - \frac{AB_t p_t}{(b_t^T p_t)^2} \left(\frac{\partial b_t^T}{\partial \theta_j} p_t \right)$$

d'où :

$$\frac{\partial p_{t+1}}{\partial \theta_j} = \frac{AB_t}{b_t^T p_t} \left[I - \frac{p_t b_t^T}{b_t^T p_t} \right] \frac{\partial p_t}{\partial \theta_j} + \left(\frac{\partial A}{\partial \theta_j} B_t + A \frac{\partial B_t}{\partial \theta_j} \right) \frac{p_t}{b_t^T p_t} - \frac{AB_t p_t}{(b_t^T p_t)^2} \left(\frac{\partial b_t^T}{\partial \theta_j} p_t \right) \tag{4}$$

avec, si p_1 est la distribution initiale : $\frac{\partial p_1}{\partial \theta_j} = 0$ pour tout j.

Le reste du calcul de la dérivée ne comporte pas de difficultés. On peut donc calculer pour un coût calcul raisonnable, la log-vraisemblance et sa dérivée, ce qui permet d'utiliser une des nombreuses techniques d'optimisation différentielle pour approcher un maximum, au moins local, de la log-vraisemblance.

4 Application : Etude des pics de pollution en ozone

Le but de cette étude est de prédire le maximum journalier du taux de pollution en niveau d'ozone durant la période d'avril à septembre inclus. Pour cela on utilise comme régresseurs la maximum du taux de pollution en niveau d'ozone de la veille et les observations météorologiques suivantes :

- La radiation globale
- La vitesse moyenne du vent du jour
- La température maximale de la journée
- Le gradient de température sur un jour

La modélisation statistique de l'ozone et plus particulièrement des modèles de régression a été beaucoup étudiée. Les modèles linéaires ne semblent pas capturer toute la complexité du phénomène. C'est pourquoi il faut employer des modèles plus riches (cf Chen, Islam and Biswas [7] ou Gardner et Dorling [11]). Parmi ces modèles, les MLP semblent donner de meilleurs résultats que les modèles linéaires bien qu'ils demandent souvent beaucoup plus d'efforts pour les mettre en oeuvre pour obtenir seulement une modeste amélioration des prédictions. (cf Comrie [8]).

On montre ici comment on peut encore améliorer cette prédiction grâce au modèle hybride HMM/MLP. De plus, outre l'amélioration de la prédiction, cette modélisation apporte des informations supplémentaires précieuses pour prédire les pics de pollution.

Pour cette étude, nous disposons des observations météorologiques et du taux de pollution d'ozone pour les années de 1994 jusqu'à 1997 inclus. Nous utiliserons les données de 1994 jusqu'à 1996 pour estimer nos modèles (données "in sample"), et nous comparerons les différents modèles sur les données 1997 (données "out of sample).

4.1 Comparaison entre le MLP et le modèle linéaire

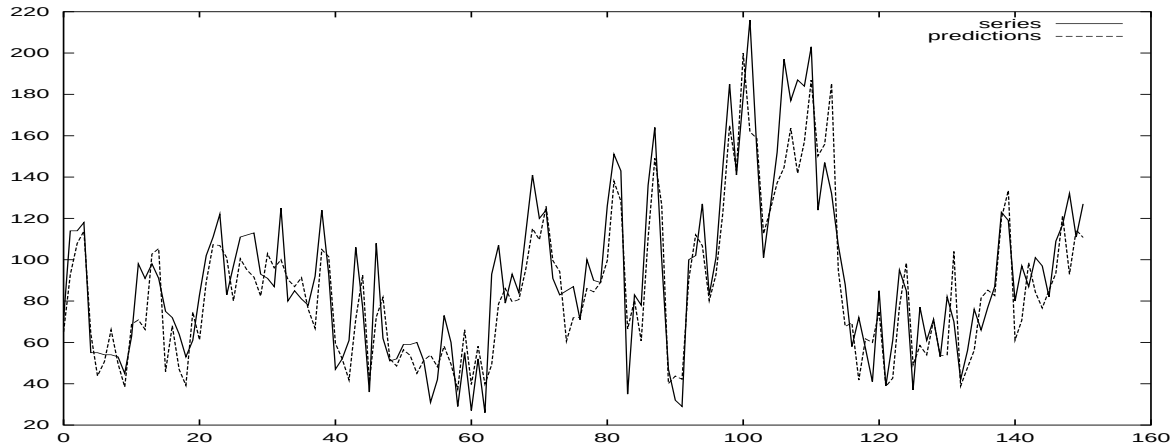
Cette étude préliminaire permet de voir les apports du MLP par rapport au modèle linéaire simple. L'architecture du modèle MLP a été déterminée grâce à la méthode SSM exposée dans la première partie. Notre critère de performance est ici la racine carrée de l'erreur quadratique moyenne (RMSE), elle est exprimée en microgramme d'ozone par mètre cube ($\mu g/m^3$). Le tableau 1 résume les résultats obtenus :

Table 1: Comparaison MLP, modèle linéaire

Années	1994-1996 (in sample)	1997 (out of sample)
RMSE MLP	17.49 $\mu g/m^3$	17.98 $\mu g/m^3$
RMSE LINEAR	20.97 $\mu g/m^3$	19.70 $\mu g/m^3$

On remarque premièrement que le MLP améliore notablement les performances du modèle linéaire, que ce soit sur les données "in sample" ou "out of sample". Bien que le nombre de données "in sample" pour estimer le modèle ne soit pas très grand (550), la méthode SSM permet d'éviter le sur-apprentissage de façon tout à fait satisfaisante, puisque la différence du RMSE entre les périodes 1994-1996 et 1997 est relativement faible. Remarquons de plus que ces résultats sont tout à fait cohérents avec les précédentes études sur la pollution atmosphérique parisienne (cf Bel et al. [6]).

FIG. 3 – Prédictions du MLP sur les données "out of sample"



La figure 3 compare la vraie valeur du taux d'ozone et sa prédiction par le MLP sur la série "out of sample". On remarque que la prédiction pour les valeurs moyennes est particulièrement bonne, cependant les pics sont généralement sous-estimés. Ce comportement est d'autant plus gênant que ce sont les fortes valeurs qui intéressent les pouvoirs publics. Nous allons donc utiliser un modèle hybride HMM/MLP en espérant qu'un expert se spécialise dans la prédiction des valeurs moyennes et faibles, alors que l'autre cherche à capturer la dynamique des fortes valeurs.

4.2 Performance du modèle hybride sur la série d'ozone

Puisque les modèles linéaires sont capables de modéliser correctement les faibles valeurs de pollutions en niveau d'ozone, nous choisissons d'utiliser pour notre modèle hybride un modèle de régression linéaire et un modèle MLP ayant pour architecture celle qui a été déterminée par l'étude de la section précédente. Ce choix est guidé par le souci de garder un nombre de paramètres le plus raisonnable possible.

Après estimation des paramètres, nous obtenons la matrice de transition suivante pour la chaîne de Markov cachée :

$$\hat{A} = \begin{pmatrix} 0.97 & 0.02 \\ 0.03 & 0.98 \end{pmatrix}$$

On remarque que les termes diagonaux, qui représente la probabilité de rester dans le même état, sont très proches de la probabilité maximale 1. Cela signifie que le modèle reste pendant de longues plages dans le même état, c'est le signe qu'il a bien identifié deux régimes distincts. Les déviations standards pour l'expert linéaire σ_0 et le MLP σ_1 sont les suivantes :

$$\begin{cases} \sigma_0 = 0.11 \\ \sigma_1 = 0.20 \end{cases}$$

Ce résultat est cohérent avec l'intuition, puisque nous verrons que le modèle linéaire s'est spécialisé dans le partie facile de la série : les valeurs moyennes ou basses, ce qui lui permet de faire de bonnes prédictions, alors que le MLP est spécialisé dans la partie difficile : les fortes valeurs.

Finalement, les résultats en terme d'erreurs de prédiction sont les suivants :

Années	1994-1996	1997
RMSE	$16.51 \mu g/m^3$	$16.75\ \mu g/m^3$

On remarque d'abord une amélioration significative de l'erreur de prévision par rapport au MLP simple. De plus le modèle hybride donne des informations encore plus riches que le modèle de régression simple. On obtient en effet une segmentation de la série suivant la probabilité conditionnelle des deux régimes (cf figure 4). On remarque que les fortes probabilités du régime associé au MLP sont plutôt pour les forte valeurs, en outre il n'y a jamais de pic de pollution lorsque que cette probabilité est faible.

On peut aussi décomposer la prédiction du modèle en deux prédictions : celle de l'expert linéaire et celle du MLP. Cela laisse plus de souplesse pour l'utilisateur. En effet si celui-ci ne s'intéresse qu'aux fortes valeurs, il ne considérera que les prédictions du MLP qui sont bien plus pertinentes. Les figures 5 et 6 montre le 'scatterplot' des prédictions des deux experts par rapport aux vraies valeurs sur l'ensemble des données (in et out of sample)

Figure 4: Série centrée normée et probabilité de l'état associé au MLP

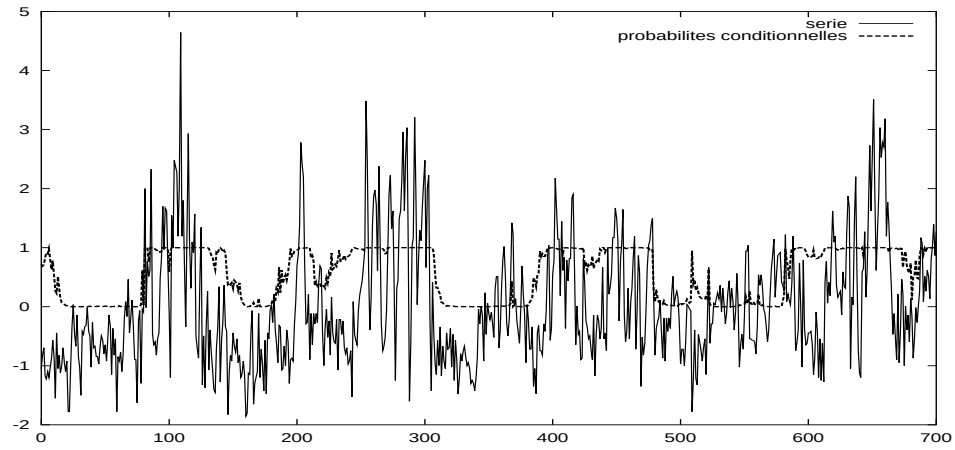
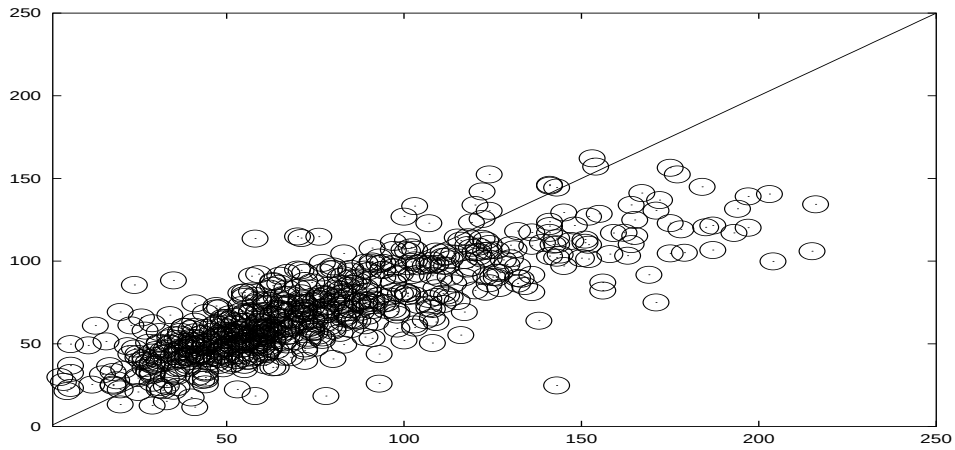


Figure 5: Prédictions de l'expert linéaire, en fonction des vraies valeurs

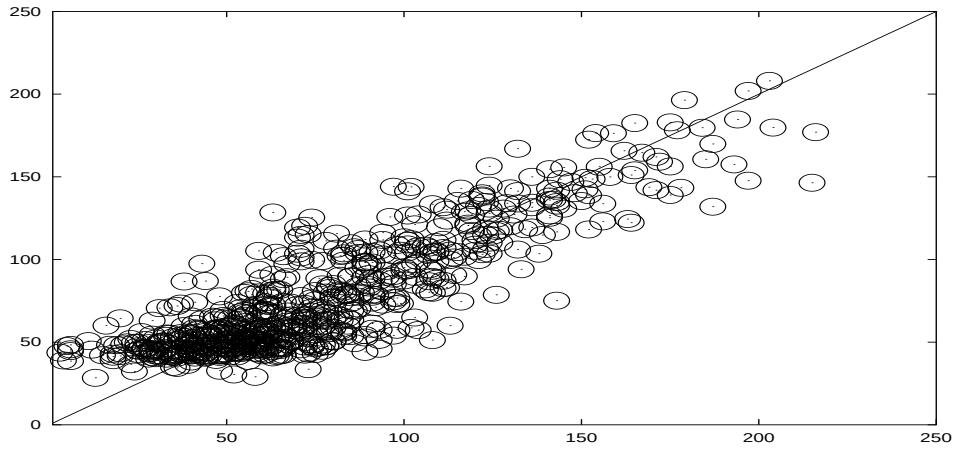


On remarque sur ces figures que l'expert linéaire est meilleur que le MLP sur les valeurs basses, cependant, pour les fortes valeurs, ses prédictions sont nettement plus petites que les vraies valeurs. Le comportement de l'expert MLP est l'opposé, il surestime les faibles valeurs mais il estime beaucoup mieux les fortes. Si on utilise uniquement ce MLP pour faire les prédictions, l'erreur quadratique moyenne sera moins bonne que le modèle autorégressif simple, mais elle sera meilleure sur la partie intéressante de la série : les pics de pollution en niveau d'ozone.

5 Conclusion

Nous avons introduit ici deux modèles importants utilisés pour des séries temporelles. La première partie a été consacré au difficile problème du dimensionnement du modèle. Nous fournissons ici une méthodologie, basée sur des propriétés asymptotiques des modèles paramétriques de régression. D'après notre expérience, celle-ci donne de bons résultats dés que l'on dispose d'un nombre suffisamment important d'observations (au moins 500).

Figure 6: Prédictions de l'expert MLP, en fonction des vraies valeurs



Dans la seconde partie, nous avons généralisé notre problème et examiné le cas des séries temporelles stationnaires par morceaux. La dynamique plus complexe de ces séries ne peut pas être capturée par un modèle de régression simple, on utilise donc, pour une seule série, plusieurs fonctions de régression qui sont articulées par une chaîne de Markov cachée. Toutes les fonctions autorégressives font simultanément une prédiction sur cette série, le rôle de la chaîne de Markov cachée est alors de pondérer les prédictions de ces modèles de régression par les probabilités conditionnelles des différents régimes. En utilisant ce modèle sur des données de la pollution en niveau d'ozone à Paris, nous avons montré que ceux-ci semblent prometteurs pour prédire des phénomènes probablement associés à des changements de régimes tels que les pics de pollution.

On doit cependant remarquer qu'il n'existe pas encore d'outils statistiques pour choisir le nombre de régimes pour une série. En effet si on surestime le nombre de régimes, on perd l'identifiabilité du modèle, et les outils statistiques classiques tels que ceux utilisés dans la première partie pour déterminer l'architecture des MLP, ne sont plus justifiables théoriquement. Ce problème est d'ailleurs un domaine de recherche très actif, qui utilise des outils statistiques et mathématiques très complexes qui dépasse largement le cadre de cet article.

Références

[1] H. Akaike. A new look at the statistical model identification. *Transactions on automatic Control*, 19 :716–723, 1974.

[2] L.E. Baum. An inequality and associated maximization technique in statistical estimation for probalistic functions of Markov processes. *Inequalities*, 3 :1–8, 1972.

[3] L.E. Baum and A. Egon. An inequality with applications to statistical estimation for probalistic functions of a Markov process and to a model for ecology. *Bull. Amer. Meteorol. Soc.*, 73 :360–363, 1967.

[4] L.E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite Markov chains. *Annals of Mathematical statistics*, 37 :1559–1563, 1966.

[5] L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximisation technique occuring in the statistical estimation of probabilistic functions of Markov processes. *Annals of Mathematical statistics*, 41 :1 :164–171, 1970.

[6] L. Bel et. al. Elément de comparaison de prévisions statistiques des pics d'ozone. *Revue de Statistique appliquée*, 47 :3 :7–25, 1972.

[7] J.L. Chen, S. Islam, and P. Biswas. Nonlinear dynamics of hourly ozone concentrations : nonparametric short-term prediction. *Atmospheric Environment*, 32 :1839–1848, 1998.

[8] A.C. Comrie. Comparing neural network and regression models for ozone forecasting. *Journal of the Air and Waste Management Association*, 47 :653–663, 1997.

[9] M. Cottrell, et al. Neural modeling for time series : a statistical stepwise method for weight elimination. *IEEE Transaction on Neural Networks*, 6 :1355–1364, 1995.

[10] R. Douc, E. Moulines, and T. Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regimes. Technical reports 9, University of Lund, 2001.

[11] M.W. Gardner and S.R. Dorling. Statistical surface ozone models : an improved methodology to account for non-lionear behaviour. *Atmospheric environment*, 34 :21–34, 2000.

[12] J. D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57 :357–384, 1989.

[13] V. Krishnamurthy and T. Rydén. Consistent estimation of linear and non-linear autoregressive models with Markov regime. *Journal of time series analysis*, 19 :3 :291–307, 1998.

[14] F. LeGLand and L. Mevel. Exponential forgetting and geometric ergodicity in Hidden Markov Models. *Mathematics of control, Signal, and Systems, à paraître*, 1999.

[15] J. Rynkiewicz. Modèles hybrides intégrant des réseaux de neurones artificiels à des modèles de chaînes de Markov cachées : applications à la prediction de séries temporelles. PhD thesis, Université de Paris 1, 2000.

[16] J. Rynkiewicz. Estimation of Hybrid HMM/MLP models. In *ESANN'2001*, 2001.

[17] G. Schwarz. estimating the dimension of a model. *The Annals of Statistics*, 6 :2 :461–464, 1978.

[18] H.J. Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output Map. *Neural Networks*, 5 :589–593, 1992.

[19] J. Yao. On least square estimation for stable nonlinear AR processes. *The Annals of Institut of Mathematical Statistics*, 52 :316–331, 2000.