

REVUE DE STATISTIQUE APPLIQUÉE

W. IMAM

E. MATZNER-LØBER

A.-S. AIFOUTE

Choix de l'ordre des modèles autorégressifs fonctionnels additifs

Revue de statistique appliquée, tome 47, n° 1 (1999), p. 63-80

http://www.numdam.org/item?id=RSA_1999__47_1_63_0

© Société française de statistique, 1999, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

CHOIX DE L'ORDRE DES MODÈLES AUTORÉGRESSIFS FONCTIONNELS ADDITIFS

W. Imam¹, E. Matzner-Løber² et A.-S. Aifoute³

¹ Université de Damas, Faculté de génie électrique et mécanique,

Dép^t de génie médical, Damas, Syrie

² Unité de Biométrie, ENSA.M, INRA, Université Montpellier II,

2 Place Viala, 34060 Montpellier Cedex 1, France

Téléphone : +33-99-61-2422

³ Faculté de Médecine, Université de Montpellier I,

2, rue Ecole de médecine, 34060 Montpellier Cedex 1, France

RÉSUMÉ

Dans cet article, nous considérons une classe de modèles additifs non linéaires avec variables exogènes dans le cadre de l'analyse non linéaire des séries chronologiques. Nous proposons deux procédures pour caractériser ces modèles. La première utilise une méthode d'analyse des données, l'analyse en composantes principales avec variables instrumentales (ACPVI) spline additive. La seconde utilise les fonctions splines de régression. Après avoir déterminé l'ordre du modèle, nous sélectionnons les variables influentes et nous identifions les fonctions, linéaires ou non, impliquées. Nous présentons ensuite des résultats de simulations ainsi qu'un exemple pour illustrer cette analyse. Pour terminer, nous appliquons la première procédure à un exemple médical concret.

Mots-clés : ACPVI spline additive, Choix de modèle, Identification, Séries chronologiques.

ABSTRACT

In this article, we consider a class of nonlinear additive models with exogenous variables for nonlinear times series analysis and propose two procedures for detecting such models. The first one uses a data analysis method called additive spline principal components analysis with respect to instrumental variables (PCAIV). The second one uses the regression splines. Once the variable selection is done, we identify the linear and nonlinear functions involved. Hence, we present the results of a simulation study and use a simulated example to illustrate the analysis. Then, we use the first procedure in a real medical example.

Keywords : Additive spline, PCAIV, Time series, Variable selection.

1. Introduction

Ces dernières années, l'étude non linéaire des séries chronologiques a connu un essor important. Cet engouement peut s'expliquer par diverses raisons, comme par

exemple la limitation des modélisations linéaires dans un grand nombre d'applications réelles ou encore l'apparition d'ordinateurs puissants qui permettent d'effectuer rapidement les nombreux calculs nécessaires à l'analyse non linéaire de ces séries. Une troisième raison à cet essor est le développement des méthodes de régression non paramétriques qui ont donné une base théorique solide à l'étude non linéaire des séries temporelles.

Dans ce travail, nous nous intéressons plus particulièrement aux méthodes non paramétriques utilisées dans l'étude des séries chronologiques. Il existe de nombreux modèles non linéaires dans la littérature (bilinéaire, ARCH, EXPAR...), mais ce n'est pas l'objet de cet article de les détailler et de les étudier. Nous renvoyons les lecteurs intéressés à Tong (1995). Les premières applications des méthodes non paramétriques aux séries chronologiques étaient consacrées à la prévision. Elles sont dues à Bosq (1979) et Collomb (1980) qui ont étudié le régressogramme. Ils ont été suivis par Collomb, Härdle et Hassani (1987) (étude du mode conditionnel), Gannoun (1989) (étude de la médiane conditionnelle). Les prédicteurs non paramétriques basés sur l'estimation non paramétrique de la moyenne, de la médiane et du mode conditionnels ont ensuite été repris sous diverses formes et étudiés sous différentes hypothèses de mélange. Les premières applications supposaient la série k -Markovienne, c'est-à-dire des modèles de la forme suivante

$$z_t = f(z_{t-1}, \dots, z_{t-k}) + \varepsilon_t, \quad (1)$$

où (ε_t) est une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d.), de moyenne 0 et de variance σ^2 , indépendante du passé du processus. Les modèles (1) sont des processus autorégressifs non linéaires (ANL) d'ordre k .

Pour ce qui est de l'analyse des séries elle-même, Robinson (1983) a étudié des estimateurs non paramétriques de la densité jointe de la série stationnaire Z_t pour différentes valeurs de t , des densités conditionnelles et également des courbes de régression. Considérant cette approche non paramétrique pour l'analyse des séries chronologiques, il proposait ces méthodes afin d'analyser les distributions des processus stationnaires. Ainsi dans son introduction, Robinson signalait que ces méthodes pouvaient permettre de dégager la nature gaussienne ou non de la série chronologique étudiée, ou encore de détecter son caractère linéaire ou non linéaire.

Ce n'est qu'en 1990 que Auestad et Tjøstheim ont appliqué les méthodes non paramétriques de régression pour estimer la moyenne et la variance conditionnelles d'un processus autorégressif non linéaire. Ils ont également proposé quelques pistes pour déterminer l'ordre k d'un ANL. La première consistait à remarquer que l'estimateur de la variance résiduelle d'ordre j (j variant de 1 à une valeur prédéterminée K) devait se stabiliser pour j plus grand que k . L'estimateur de la variance résiduelle $\hat{\sigma}_j^2$ était de la forme suivante :

$$\hat{\sigma}_{1, \dots, j}^2 = \frac{1}{(n-j)} \sum_{t=j+1}^n (z_t - r_n(z_{t-1}, \dots, z_{t-j}))^2,$$

où r_n est un estimateur non paramétrique de la régression. Ils proposaient également un estimateur du critère FPE (Final Prediction Error) pour déterminer cet ordre. Ils ont

ensuite repris cette approche en 1994 (Tjøstheim et Auestad, 1994b). Entre temps, Cheng et Tong (1992) ont proposé d'utiliser un critère de validation croisée. Vieu (1995) a également proposé un critère de validation croisée pénalisée. Ses simulations confirment les résultats théoriques qu'il a obtenus.

Comme l'approche non paramétrique en général souffre de la «malédiction de la dimension», même quand k est petit, des restrictions sur la fonction f ont été proposées. La plus utilisée est l'additivité; le modèle (1) peut alors s'écrire sous la forme suivante

$$z_t = f_1(z_{t-1}) + f_2(z_{t-2}) + \cdots + f_k(z_{t-k}) + \varepsilon_t, \quad (2)$$

où (ε_t) est une suite de variables aléatoires i.i.d., les $f_i(\cdot)$ sont des fonctions réelles appelées fonctions coordonnées. Ces ANL additifs seront dorénavant notés ANLA ou NAAR en anglais. Les avantages de cette restriction sont nombreux : meilleure vitesse de convergence, facilité d'interprétation... En régression, les modèles additifs ont été très étudiés (voir par exemple Hastie et Tibshirani, 1990). Les ANLA sont une simple généralisation des modèles autorégressifs non linéaires de premier ordre de Jones (1978). Ils sont suffisamment flexibles pour englober des modèles autorégressifs linéaires mais également beaucoup de modèles non linéaires intéressants. Pour une présentation plus approfondie des ANLA, nous renvoyons à l'article de Chen et Tsay (1993). Si une variable exogène X est disponible, le modèle (2) peut se généraliser de la façon suivante

$$z_t = f_1(z_{t-1}) + f_2(z_{t-2}) + \cdots + f_k(z_{t-k}) + g_1(x_{t-1}) + \cdots + g_l(x_{t-l}) + \varepsilon_t, \quad (3)$$

où les $g_i(\cdot)$ sont des fonctions réelles mesurables, l est un entier positif et les variables exogènes x_t sont indépendantes du bruit de la série (ε_t) .

Chen et Tsay (1993) ont utilisé un «Backfitting algorithm» comme l'Alternating Conditional Expectation (ACE) introduit par Breiman et Friedman (1985) et le BRUTO algorithme de Hastie et Tibshirani (1990), pour ajuster ces modèles additifs (3). L'algorithme AVAS de Tibshirani (1988) pourrait également être utilisé. Le problème rencontré par Chen et Tsay (1993) est un problème inhérent aux algorithmes de backfitting, c'est-à-dire que lorsque les observations sont très corrélées, la vitesse de convergence des algorithmes est très lente.

Une autre méthode pour estimer les fonctions dans les ANLA est de considérer un estimateur par intégration. Nous renvoyons pour de plus amples détails aux articles de Tjøstheim et Auestad (1994a) ou encore de Masry et Tjøstheim (1995).

Dans ce travail, nous allons utiliser une méthode d'analyse des données, l'analyse en composantes principales sur variables instrumentales spline additive (ACPVI spline additive), introduite par Durand (1993), pour déterminer l'ordre d'un ANLA. L'ACPVI spline additive et les splines de régression sont présentées dans les sections 2 et 3. La section 4 présente la procédure d'estimation de l'ordre des ANLA en utilisant l'ACPVI spline additive ainsi que les résultats des simulations. Pour terminer ce travail, nous présentons une étude concernant un problème médical.

2. Rappels sur l'ACPVI spline additive

L'analyse linéaire en composantes principales sur variables instrumentales (ACPVI) est une méthode ayant pour objectif l'analyse des relations linéaires entre deux groupes de variables \mathbf{X} et \mathbf{Y} . \mathbf{X} , matrice de taille $(n \times p)$, représente les mesures de p variables explicatives sur n individus et \mathbf{Y} , matrice de taille $(n \times q)$, représente les mesures de q variables à expliquer sur les mêmes n individus. Cette méthode a été introduite pour la première fois par Rao (1964). L'approche proposée par Escoufier (1987) réside en l'introduction des métriques (matrices définies positives) permettant de calculer les distances entre individus et entre variables.

Les métriques \mathbf{Q} et \mathbf{D} sont respectivement définies positives sur les espaces des individus et des variables de \mathbf{Y} . L'objectif de l'ACPVI est de trouver une métrique $\bar{\mathbf{R}}$ sur l'espace des individus de \mathbf{X} telle que la norme de Hilbert Schmidt, (norme sur l'espace des matrices D-symétriques) de la différence entre les opérateurs $\mathbf{YQY}'\mathbf{D}$ et $\mathbf{X}\bar{\mathbf{R}}\mathbf{X}'\mathbf{D}$, soit minimale. Cela revient à chercher des combinaisons linéaires des variables explicatives qui expliquent le mieux les variables réponses dans l'espace engendré par les variables explicatives.

2.1. Transformations splines des prédicteurs

L'hypothèse implicite de liaison linéaire entre les variables explicatives n'est pas toujours vérifiée dans de nombreuses situations biologique, médicale et écologique. L'ACPVI ne prend pas en compte le caractère non linéaire des liaisons entre les variables explicatives, Durand (1993) a proposé l'ACPVI spline additive. Cette méthode cherche des transformations des variables explicatives par des fonctions splines pour améliorer la corrélation entre les variables explicatives et les variables réponses. Ces fonctions sont des polynômes par morceaux d'ordre m (de degré $m-1$) qui se raccordent ainsi que certaines de leurs dérivées en h points appelés nœuds intérieurs, voir Schumaker, (1981). L'espace des fonctions splines est de dimension $r = m + h$. Une fonction spline $s(z)$ est une combinaison linéaire de fonctions de base appelées B -splines notées $B_l(\cdot)$ où l varie de 1 à r . La fonction spline s'écrit de la façon suivante :

$$s(z) = \sum_{l=1}^r s_l B_l(z),$$

où $z \in \mathbb{R}$ et les s_l sont les coefficients splines. Chaque élément x_i^j de la $j^{\text{ème}}$ variable explicative ($j^{\text{ème}}$ colonne de \mathbf{X}) est codé par $(B_1(x_i^j), B_2(x_i^j), \dots, B_r(x_i^j))$. La transformation de la $j^{\text{ème}}$ variable explicative \mathbf{X}^j s'écrit de la façon suivante :

$$\begin{aligned} \mathbf{X}_{(n \times 1)}^j(\mathbf{s}^j) &= \left[\sum_{l=1}^r B_l(x_1^j) s_l^j, \dots, \sum_{l=1}^r B_l(x_n^j) s_l^j \right]' \\ &= \mathbf{B}_{(n \times r)}^j \mathbf{s}_{(r \times 1)}^j. \end{aligned}$$

où \mathbf{B}^j est la matrice de codage de \mathbf{X}^j et $\mathbf{s}^j = [s_1^j, \dots, s_r^j]'$ le vecteur des coefficients splines. Ces coefficients seront estimés en minimisant une fonction objectif détaillée plus loin.

2.2. La méthode

L'Analyse en Composantes Principales (ACP) a été utilisée par Ciuperca (1997) sur les variables exogènes du modèle (3) pour rechercher les variables influant sur la variable réponse (la série z_t) qui est considérée comme une variable supplémentaire. Cette méthode ne tient pas compte de la corrélation entre les variables exogènes et la (les) variable(s) réponse(s) et de la possible non-linéarité des variables exogènes.

L'objectif de l'ACPVI spline additive est de chercher des combinaisons linéaires des variables instrumentales (explicatives) transformées par des fonctions splines qui expliquent le mieux les variables réponses dans l'espace engendré par les variables instrumentales transformées. Cela revient à effectuer une analyse en composantes principales de la projection de \mathbf{Y} dans l'espace engendré par les variables instrumentales transformées par les splines.

L'ACPVI spline additive cherche une matrice $\mathbf{X}(\bar{\mathbf{s}})$ et une métrique $\bar{\mathbf{R}}$ qui minimisent la distance entre les opérateurs $\mathbf{YQY}'\mathbf{D}$ et $\mathbf{X}(\mathbf{s})\mathbf{R}\mathbf{X}'(\mathbf{s})\mathbf{D}$.

$$\begin{aligned} (\bar{\mathbf{R}}, \bar{\mathbf{s}}) &= \arg \min_{(\mathbf{R}, \mathbf{s})} \|\mathbf{YQY}'\mathbf{D} - \mathbf{X}(\mathbf{s})\mathbf{R}\mathbf{X}'(\mathbf{s})\mathbf{D}\|^2 \\ &= \arg \min_{(\mathbf{R}, \mathbf{s})} \text{tr}[(\mathbf{YQY}'\mathbf{D} - \mathbf{X}(\mathbf{s})\mathbf{R}\mathbf{X}'(\mathbf{s})\mathbf{D})^2], \end{aligned} \tag{4}$$

où $\mathbf{X}_{(n \times p)}(\mathbf{S}) = \mathbf{B}_{(n \times pr)}\mathbf{S}_{(pr \times p)}$ avec $\mathbf{B}_{(n \times pr)} = [\mathbf{B}^1 | \dots | \mathbf{B}^p]$ la matrice du codage, $\mathbf{B}_{(n \times r)}^j$ étant, rappelons-le, la matrice du codage de la $j^{\text{ème}}$ variable explicative ($1 \leq j \leq p$), et

$$\mathbf{S}_{(pr \times p)} = \begin{bmatrix} s_1^1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ s_r^1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & s_1^p \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & s_r^p \end{bmatrix},$$

est la matrice des coefficients splines, s_l^j ($1 \leq j \leq p$, $1 \leq l \leq r$) étant le coefficient de $B_l(\mathbf{X}^j)$ dans la transformation $\mathbf{X}^j(\mathbf{s}^j)$. Le vecteur $\mathbf{s} = \mathbf{S}\mathbf{1}_p$ de \mathbb{R}^{pr} , où $\mathbf{1}_p$ est le vecteur colonne de 1, contient les paramètres splines pour tous les prédicteurs. Les équations normales par rapport à \mathbf{R} , \mathbf{s} étant fixé, ont une solution explicite optimale calculée par

$$\mathbf{R} = (\mathbf{X}'(\mathbf{s})\mathbf{D}\mathbf{X}(\mathbf{s}))^+ (\mathbf{X}'(\mathbf{s})\mathbf{D}\mathbf{YQY}'\mathbf{D}\mathbf{X}(\mathbf{s})) (\mathbf{X}'(\mathbf{s})\mathbf{D}\mathbf{X}(\mathbf{s}))^+.$$

Pour \mathbf{R} fixé, les équations normales n'ont pas de solution explicite; pour obtenir s , il faut donc utiliser une méthode itérative comme celle proposée par Durand, (1993) et Imam et Durand, (1997). La condition $n \gg p$ est importante pour ne pas avoir un problème de sensibilité numérique des algorithmes utilisés.

Une fois la transformation optimale effectuée, nous réalisons l'ACP du triplet $(\mathbf{X}(\bar{s}), \bar{\mathbf{R}}, \mathbf{D})$. L'ACP considérée est équivalente à celle du triplet $(\mathbf{P}_{\mathbf{X}(\bar{s})} \mathbf{Y}, \mathbf{Q}, \mathbf{D})$, où $\mathbf{P}_{\mathbf{X}(\bar{s})}$ est le projecteur D -orthogonal sur l'espace des colonnes de $\mathbf{X}(\bar{s})$. Ceci met en évidence le lien avec les lisseurs additifs multiréponses

$$\widehat{\mathbf{Y}}_{(n \times q)} = \mathbf{P}_{\mathbf{X}(\bar{s})} \mathbf{Y} = \mathbf{X}_{(n \times p)}(\bar{s}) \bar{\mathbf{M}}_{(p \times q)} = \mathbf{B}_{(n \times pr)} \bar{\mathbf{S}}_{(pr \times p)} \bar{\mathbf{M}}_{(p \times q)}, \quad (5)$$

où $\bar{\mathbf{M}} = (\bar{\mathbf{S}}' \mathbf{B}' \mathbf{D} \mathbf{B} \bar{\mathbf{S}})^+ \bar{\mathbf{S}}' \mathbf{B}' \mathbf{D} \mathbf{Y}$ est la matrice des coefficients de la régression de \mathbf{Y} sur la matrice $\mathbf{X}(\bar{s})$. Le modèle additif multiréponses associé à l'ACPVI spline additive s'exprime par

$$\widehat{\mathbf{Y}}^h = \sum_{j=1}^p \mathbf{B}^j \bar{s}^j \bar{\mathbf{M}}_{jh}, \quad h = 1, \dots, q. \quad (6)$$

On peut alors, pour chaque variable réponse \mathbf{Y}^h , tracer $\mathbf{B}^j \bar{s}^j \bar{\mathbf{M}}_{jh}$ en fonction de \mathbf{X}^j , ce qui conduit aux fonctions coordonnées reconstruisant le modèle additif. La variance des valeurs d'une fonction coordonnée est un critère permettant de juger de l'influence de cette variable explicative sur la variable réponse.

Dans la suite de ce travail, nous ne considérerons que des séries temporelles univariées, \mathbf{Y} sera donc de taille $(n \times 1)$. Nous expliquerons en détail au paragraphe 4, la manière de construire \mathbf{X} et \mathbf{Y} à partir d'une série chronologique. Dans cette configuration, l'ACPVI spline additive entre \mathbf{X} et \mathbf{Y} revient à projeter la réponse \mathbf{Y} dans l'espace engendré par les colonnes de $\mathbf{X}(\bar{s})$. Les équations (5) et (6) deviennent

$$\widehat{\mathbf{Y}}_{(n \times 1)} = \mathbf{B}_{(n \times pr)} \bar{\mathbf{S}}_{(pr \times p)} \bar{\mathbf{M}}_{(p \times 1)} = \sum_{j=1}^p \mathbf{B}^j \bar{s}^j \bar{\mathbf{M}}_j, \quad (7)$$

où $\bar{\mathbf{S}}$ est calculée grâce à (4).

Une autre approche dite codage, permettant de considérer l'influence des variables \mathbf{X} sur \mathbf{Y} , utilise les splines de régression. Cette méthode cherche à expliquer la réponse \mathbf{Y} dans l'espace engendré par les colonnes de la matrice du codage \mathbf{B} .

3. Rappels sur les splines de régression

Le principe de cette méthode consiste à estimer les fonctions f_j du modèle choisi par des fonctions splines. Les fonctions f_j s'écrivent comme une somme des fonctions splines exprimées dans la base des B-splines normalisées. Pour des nœuds

fixés, on obtient

$$\begin{aligned}
 y_i &= \sum_{j=1}^p f_j(x_i^j) + \varepsilon_i \\
 &= \sum_{j=1}^p \sum_{l=1}^r \alpha_l^j B_l^j(x_i^j) + \varepsilon_i \quad \text{pour } i = 1, \dots, n.
 \end{aligned}$$

Le modèle des splines de régression est donné par l'expression

$$\mathbf{Y}_{(n \times 1)} = \mathbf{B}_{(n \times pr)} \mathbf{A}_{(pr \times 1)} + \boldsymbol{\varepsilon}_{(n \times 1)}, \tag{8}$$

où $\boldsymbol{\varepsilon}' = [\varepsilon_1, \dots, \varepsilon_n]$, \mathbf{A} est le vecteur des paramètres splines et \mathbf{B} la matrice du codage de dimension $n \times pr$. L'estimation des paramètres splines repose sur la minimisation des moindres carrés

$$\frac{1}{n} \sum_{i=1}^n \left[y_i - \sum_{j=1}^p \sum_{l=1}^r \alpha_l^j B_l^j(x_i^j) \right]^2.$$

La solution de ce problème d'optimisation est la projection de \mathbf{Y} sur l'espace engendré par les colonnes de la matrice du codage \mathbf{B} ,

$$\widehat{\mathbf{Y}} = \mathbf{P}_B \mathbf{Y} = \mathbf{B} \overline{\mathbf{M}}, \tag{9}$$

où $\overline{\mathbf{M}} = (\mathbf{B}' \mathbf{D} \mathbf{B})^+ \mathbf{B}' \mathbf{D} \mathbf{Y}$.

Dans le cas où $q = 1$, la différence entre l'ACPVI spline additive et les splines de régression provient de la matrice \mathbf{S} . La condition $n \gg pr$ est plus contraignante que celle imposée par l'ACPVI spline additive. D'un point de vue algébrique, cette condition permet d'éviter les problèmes de sensibilité numérique des algorithmes utilisés. D'un point de vue statistique, pr est la dimension de l'espace dans laquelle on projette la réponse de taille n . Si cet espace est trop grand, le modèle possède alors un mauvais pouvoir prédictif. Nous verrons dans la partie simulations que cette condition n'est pas respectée pour les séries de taille $T = 50$. Même pour des séries de taille supérieure, cette condition n'est pas respectée lorsque le nombre p de variables explicatives est élevé. C'est une des raisons pour laquelle les résultats des splines de régression sont de moindre qualité que ceux de l'ACPVI spline additive.

4. Méthodologie et simulations

Soit (Z_t) , $t \in \mathbb{N}$, un processus ANLA univarié, dont on observe un échantillon de taille T . À partir de ce processus, nous construisons un processus associé de la façon suivante :

$$\mathbf{X} = [Z_{t-1}, \dots, Z_{t-K}] \quad \text{et} \quad \mathbf{Y} = [Z_t],$$

où K est assez grand pour détecter des phénomènes saisonniers éloignés, mais pas trop afin de conserver suffisamment de données pour pouvoir étudier la série. Afin de respecter les notations des paragraphes précédents, posons $n = T - K$, les matrices \mathbf{X} et \mathbf{Y} sont donc de taille respective $n \times K$ et $n \times 1$.

La première étape consiste à déterminer un estimateur \hat{k} de la valeur k du modèle (2). Nous effectuons une ACPVI spline additive entre \mathbf{X} et \mathbf{Y} définies ci-dessus. Les résultats de cette ACPVI spline additive sont résumés de la manière suivante :

- La variance résiduelle pour chaque variable, $\hat{\sigma}_j^2 = \frac{1}{n-1} \sum_{t=K+1}^T (z_t - \hat{f}_j(z_{t-j}))^2$;

ces premiers résultats donnent une idée de l'importance des variables explicatives. $\hat{f}_j(z_{t-j})$ est soit la transformée de la variable \mathbf{X}^j dans le cas de l'ACPVI spline additive, soit estimée par des splines dans le cas des splines de régression.

- La variance résiduelle d'ordre j , $\hat{\sigma}_{1,\dots,j}^2 = \frac{1}{n-1} \sum_{t=K+1}^T (z_t - \hat{f}_1(z_{t-1}) - \dots -$

$\hat{f}_j(z_{t-j}))^2$, c'est-à-dire que l'on prend en compte toutes les variables jusqu'à la $j^{\text{ème}}$. Cette variance résiduelle d'ordre j , donne une indication de l'ordre k de l'ANLA. Nous cherchons à la manière d'Auestad et Tjøstheim (1990), la valeur de \hat{k} ($\hat{k} \leq K$) pour laquelle cette variance se stabilise. On calcule la variance résiduelle $\hat{\sigma}_{1,\dots,j}^2$ pour j variant de 1 à K et on choisit la valeur \hat{k} telle que pour $l > \hat{k}$ la variance résiduelle ne diminue pas ou peu.

Exemple : Considérons le modèle autorégressif non linéaire suivant

$$z_t = 1.5 \sin\left(\frac{\pi}{2} z_{t-2}\right) - \sin\left(\frac{\pi}{2} z_{t-3}\right) + \varepsilon_t \quad \text{où} \quad \varepsilon_t \sim \mathcal{N}(0, 1) \quad \text{et} \quad T = 50.$$

Nous choisissons $K = 5$, n vaut 45. Nous appliquons l'ACPVI spline additive entre $\mathbf{X} = [Z_{t-1}, \dots, Z_{t-5}]$ et $\mathbf{Y} = [Z_t]$. Les résultats se présentent de la façon suivante :

j	1	2	3	4	5
$\hat{\sigma}_j^2$	1.69	0.52	1.18	1.73	1.55
$\hat{\sigma}_{1,\dots,j}^2$	1.69	0.52	0.18	0.17	0.13

On remarque que la deuxième variable est très influente. Pour des valeurs de $j > 3$, nous remarquons également que la variance résiduelle $\hat{\sigma}_{1,\dots,j}^2$ ne varie presque plus à partir de $j > 3$ et nous choisissons donc $\hat{k} = 3$.

Dans un deuxième temps, nous proposons une procédure d'identification du modèle sous-jacent. Nous refaisons une ACPVI spline additive avec \hat{k} variables, où \hat{k} a été sélectionné lors de la première phase (dans la partie simulations \hat{k} pourra donc être la vraie valeur du modèle simulé mais également une valeur erronée), c'est-à-dire une ACPVI spline additive entre $\mathbf{X} = [Z_{t-1}, \dots, Z_{t-\hat{k}}]$ et $\mathbf{Y} = [Z_t]$ (en général $\hat{k} < K$, si $\hat{k} = K$, on peut penser que la valeur de K initialement choisie est trop

En utilisant ces derniers critères, nous retrouvons le vrai modèle mais pour ce faire nous avons effectué 12 ACPVI spline additive.

Pour terminer cette phase d'identification, nous réalisons une ACPVI spline additive avec les variables sélectionnées (2, 3) et nous représentons graphiquement les fonctions $\hat{f}_j(z_{t-j})$ à partir du modèle (7). Cette représentation graphique est une bonne indication de la vraie fonction, nous obtenons la figure 1.

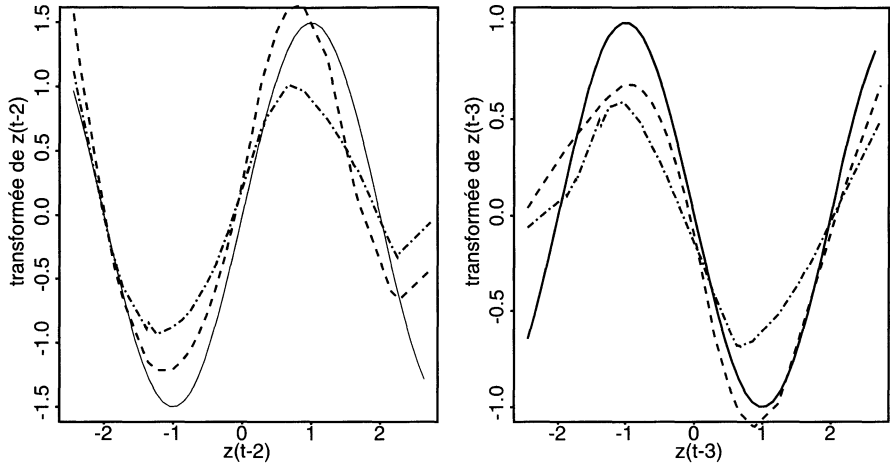


FIGURE 1

$\hat{f}_2(z_{t-2})$ et $\hat{f}_3(z_{t-3})$ obtenues avec 50 données.

Les courbes en trait plein sont les vraies fonctions, c'est-à-dire $1.5 \sin(x\pi/2)$ et $-\sin(x\pi/2)$ où x prend ses valeurs entre le minimum et le maximum de z_t . Les fonctions coordonnées estimées, calculées par la formule (7) sont tracées en pointillé. Le tracé point-tiré représente les résultats obtenus par la méthode ACE de Breiman-Friedman évoquée à la fin de l'introduction. Les résultats obtenus par l'ACPVI spline additive sont plus proches des vraies courbes que celles tracées avec la méthode ACE. Les fonctions coordonnées estimées détectent les deux sinusoides. Ainsi, avec 50 observations, l'ACPVI spline additive permet d'identifier le modèle non linéaire qui a généré la série. Nous avons suivi les mêmes démarches pour estimer l'ordre et le modèle par les splines de régression, mais, comme nous le verrons dans la partie suivante, si la condition $n \gg pr$ n'est pas respectée les résultats sont mauvais, (dans cet exemple $n = 45$ et $pr = 25$).

5. Simulations

Nous avons décidé de tester ces nouvelles procédures sur 10 séries simulées, la longueur de ces séries variant de 50 à 200 données. Nous avons considéré 5 séries

linéaires et 5 séries non linéaires définies comme suit :

$$\begin{aligned}
 S_1 \quad z_t &= 0.8z_{t-1} + \varepsilon_t, \\
 S_2 \quad z_t &= 0.8z_{t-1} + 0.6z_{t-2} + \varepsilon_t, \\
 S_3 \quad z_t &= 0.8z_{t-1} + 0.6z_{t-3} + \varepsilon_t, \\
 S_4 \quad z_t &= 0.2z_{t-1} - 0.5z_{t-3} + \varepsilon_t, \\
 S_5 \quad z_t &= 0.2z_{t-1} - 0.2z_{t-3} + \varepsilon_t, \\
 S_6 \quad z_t &= 0.8 \log(1 + 3z_{t-1}^2) + 0.6 \log(1 + 3z_{t-3}^2) + \varepsilon_t, \\
 S_7 \quad z_t &= 1.5 \sin((\pi/2)z_{t-2}) - \sin((\pi/2)z_{t-3}) + \varepsilon_t, \\
 S_8 \quad z_t &= 0.3 \sin((\pi/2)z_{t-2}) - 0.4 \sin((\pi/2)z_{t-3}) + \varepsilon_t, \\
 S_9 \quad z_t &= (0.5 - 1.1 \exp(-50z_{t-1}^2))z_{t-1} + (0.3 - 0.5 \exp(-50z_{t-3}^2))z_{t-3} + \varepsilon_t, \\
 S_{10} \quad z_t &= (0.8 - 1.1 \exp(-50z_{t-1}^2))z_{t-1} + \varepsilon_t,
 \end{aligned}$$

où (ε_t) est une suite de variables aléatoires i.i.d. de loi normale centrée réduite.

Pour chaque modèle, nous avons généré 20 réalisations de taille 50, 75, 100 et 200. Pour pouvoir appliquer la première étape de la méthode, il faut déterminer K ; il doit être suffisamment grand pour détecter des variables éloignées, mais pas trop afin de conserver assez de données pour pouvoir étudier la série. Nous avons donc choisi $K=5$ (respectivement 7, 10, 20) pour les séries de taille 50 (respectivement 75, 100, 200).

Dans un premier temps, nous avons effectué une ACPVI spline additive et les splines de régression. Pour les valeurs élevées de K (10, 20), nous avons remarqué très fréquemment que les dernières variables $(z_{t-10}, \dots, z_{t-20}$ ou $\mathbf{X}^{10}, \dots, \mathbf{X}^{20}$) n'avaient pas d'influence. Nous refaisons alors une ACPVI spline additive avec K plus petit valant entre 5 et 10. Il n'en est pas de même avec les splines de régression. Ainsi, si K vaut 10 ou 20, les résultats sont très mauvais; on ne retrouve jamais le modèle simulé.

Les tableaux 2 et 3 suivants récapitulent les résultats obtenus par les procédures que nous avons indiquées. À la fin de la première étape, on détermine une valeur pour \hat{k} . Nous comparons cette valeur à la vraie valeur du modèle. La première colonne indique le nombre de fois où la méthode utilisée détecte le bon ordre de la série sur les 20 réalisations. Une fois l'ordre sélectionné, nous appliquons l'étape 2 pour déterminer le modèle (résultats donnés dans la deuxième colonne). Nous effectuons donc une nouvelle ACPVI spline additive et une régression spline avec la valeur de \hat{k} sélectionné à la phase précédente.

Si la valeur sélectionnée est plus petite que la vraie valeur de k , nous ne pouvons en aucun cas retrouver le bon modèle. Par contre, si \hat{k} est plus grand que la vraie valeur, il est quand même possible de retrouver le vrai modèle. Ceci explique que nous avons parfois des meilleurs résultats dans le choix du modèle que dans la détermination de l'ordre.

Ces résultats sont intéressants de plusieurs points de vue. Ils permettent de voir le comportement des deux procédures que nous avons explicitées. Remarquons que ces deux méthodes donnent en général, pour un nombre de données supérieur à 100, des résultats sensiblement identiques, à condition que lors de l'utilisation de la régression spline, le K fixé au début de l'étude soit suffisamment petit. Cette restriction sur K ne permettra pas d'étudier avec cette méthode, des séries ayant un

TABLEAU 2
Résultats de l'ACPVI spline additive

Série	T=50		T=75		T=100		T=200	
	ordre	modèle	ordre	modèle	ordre	modèle	ordre	modèle
S_1	14	14	17	17	20	20	20	20
S_2	13	15	16	17	18	19	20	20
S_3	12	13	15	16	19	19	20	20
S_4	09	07	15	14	19	17	20	20
S_5	07	06	11	08	15	12	17	18
S_6	11	11	15	15	18	19	20	20
S_7	16	17	18	18	20	19	20	20
S_8	10	07	13	10	15	12	18	14
S_9	10	11	14	14	16	16	18	18
S_{10}	12	12	15	15	20	20	20	20
Total	114	113	149	144	180	173	193	190

TABLEAU 3
Résultats des splines de régression

Série	T=50		T=75		T=100		T=200	
	ordre	modèle	ordre	modèle	ordre	modèle	ordre	modèle
S_1	0	0	10	10	18	18	20	19
S_2	0	0	14	12	18	19	19	19
S_3	0	0	15	13	15	12	20	20
S_4	0	0	10	11	13	14	16	18
S_5	0	0	05	06	10	09	13	13
S_6	0	0	14	14	19	18	20	20
S_7	0	0	17	15	15	17	20	20
S_8	0	0	09	10	11	08	14	11
S_9	0	0	06	09	11	12	20	18
S_{10}	0	0	09	09	20	20	20	20
Total	0	0	109	105	150	147	182	178

comportement saisonnier. Ainsi par exemple, une série mensuelle (nous supposons que k vaut 12) de taille $T = 150$ ne pourra pas être étudiée avec la régression spline. Cette restriction sur K ne se retrouve pas dans l'ACPVI spline additive. Une autre différence de comportement correspond aux études concernant les séries S_4, S_5 et S_8 . Ces séries ont pour coefficients des valeurs proches de zéro, elles sont donc plus difficiles à analyser. Cependant l'ACPVI spline additive donne encore de bons résultats.

Pour des séries de taille plus petites ($T \leq 75$), les résultats obtenus par l'ACPVI spline additive sont corrects (conformes au modèle simulé) tant au niveau de l'ordre que de l'identification. *A contrario* les splines de régression ne donnent pas de résultats satisfaisants. Nous conseillons donc d'utiliser plutôt l'ACPVI spline additive que les splines de régression, surtout si le nombre d'observations est faible ou si le nombre de variables explicatives est élevé. C'est donc cette approche que nous allons utiliser dans l'étude des données des urgences de l'hôpital de Nîmes.

6. Application sur des données médicales

Le service des urgences est un service à part dans le complexe hospitalier. En effet, c'est d'abord une voie d'entrée de l'hôpital accessible 24 heures sur 24 heures à tous, permettant de répondre rapidement à un besoin urgent ou non. Ce besoin pouvant être médical, chirurgical, psychiatrique ou parfois même social. Cette spécificité explique le large éventail des pathologies trouvées ainsi que le nombre important d'admissions enregistrées. L'information recueillie pour l'ensemble de tous les patients est, par conséquent, considérable en volume. L'outil informatique pour la recueillir s'impose et cela pour plusieurs raisons. Il permet dans un premier temps de réguler heure par heure le service. Deuxièmement, le recueil minutieux de renseignements doit faciliter l'analyse a posteriori de l'activité des urgences sur une période donnée. C'est dans cet esprit que le service des urgences du centre hospitalier de Nîmes s'est informatisé depuis 1994.

Les diagnostics se divisent en plusieurs classes et chaque classe de pathologies médicales présente plusieurs sous-classes comme les maladies infectieuses, allergiques, pneumologiques, cardiologiques, dermatologiques, etc.

Avant d'étudier le **flux médical**, faisons une mise au point terminologique. Tout d'abord, le **flux** d'une pathologie donnée est défini par le nombre de personnes admises aux urgences par jour, pour cette pathologie. Ensuite le terme **médical** englobe toutes les pathologies appartenant à une spécialité médicale (par opposition à chirurgicale).

L'étude permettra de dégager un facteur influant sur le flux en question. Ce facteur devra être chiffrable. En l'occurrence, les facteurs qui sont mis à notre disposition sont les facteurs climatiques. Ils sont mesurables et relevés quotidiennement comme la température moyenne, le taux d'humidité dans l'air et la force du vent. Ces éléments, avec le jour de la semaine, sont les seules variables chiffrables que nous avons pu obtenir pour réaliser cette étude. Nous admettons que la variation de la fréquence d'une maladie au cours d'une période donnée peut être interprétée par des éléments médicaux ou non. C'est ainsi que l'on peut comprendre pour certaines maladies infectieuses une périodicité sur une année, en fonction des saisons. Nous savons aussi que certaines pathologies latentes peuvent être décompensées¹ à des dates bien précises. Le cas de certaines pneumopathies telles l'asthme et les insuffisances respiratoires chroniques est typique.

¹ Décompensation : rupture de l'équilibre des mécanismes régulateurs qui empêchait une affection de provoquer des troubles fonctionnels, métaboliques ou physiques. Décompenser : faire une décompensation.

De façon intuitive, nous savons que le climat peut avoir une influence directe ou indirecte sur le flux médical car il peut agir sur certains flux pathologiques le constituant. Le temps de latence entre le facteur causal et la maladie qu'il engendre (ou le jour de consultation aux urgences) peut-être logiquement compris. Le flux médical est mesuré sur l'année 1996, nous avons donc $T = 366$ données. Nous disposons de variables exogènes (température, vent...) malheureusement très corrélées; nous allons donc étudier un modèle de type (3). Dans une première étude, nous considérons 26 variables explicatives :

$$X = [F_{t-1}, \dots, F_{t-7}, T_t, T_{t-1}, \dots, T_{t-5}, H_t, H_{t-1}, \dots, H_{t-5}, V_t, V_{t-1}, \dots, V_{t-5}, J].$$

Nous réalisons une ACPVI spline additive entre X et $Y = [F_t]$ où T_t est la température à l'instant t , H_t l'humidité, V_t la force du vent, F_t le flux médical et J le jour de la semaine. Pour reprendre les notations précédentes, n vaut 359, K vaut 7 et posons $L = 19$ variables exogènes.

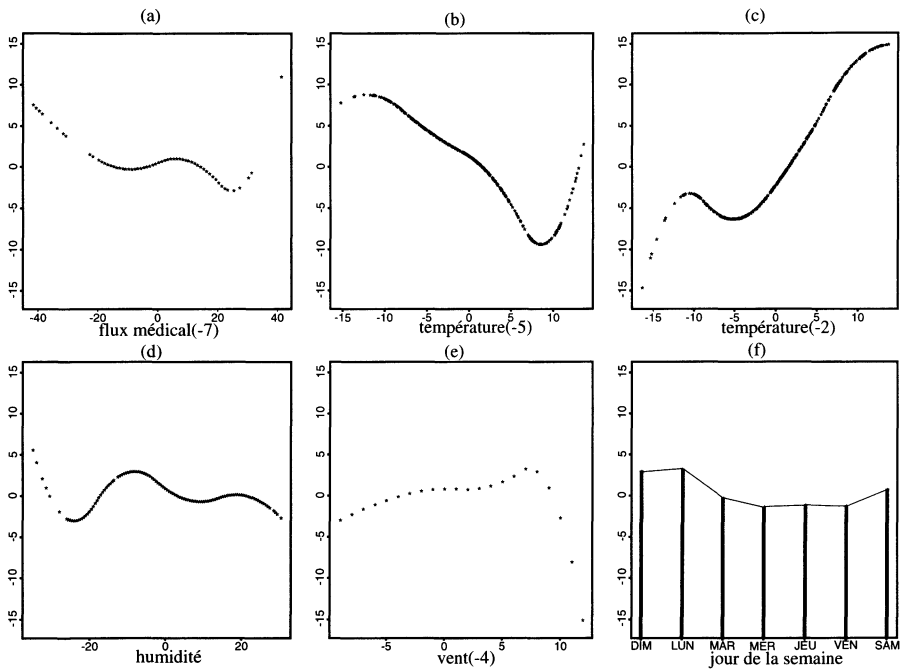


FIGURE 2

Le flux médical comme un modèle additif.

L'ACPVI spline additive a déterminé comme variables explicatives $\{F_{t-7}, T_{t-5}, T_{t-2}, H_t, V_{t-4}, J_t\}$. La représentation graphique de ces variables transformées est donnée dans la figure 2. Notons que les variables sont centrées.

Il apparaît ainsi que parmi les facteurs climatiques, la température a l'influence la plus significative sur la variation du flux des patients consultant aux urgences pour

une pathologie médicale. Ce n'est pas la température du jour de l'admission qui influe le plus, mais celles des jours qui la précèdent. Nous remarquons que plus la température est élevée, plus le flux médical est important deux jours plus tard. Alors que plus la température est basse et plus ce flux est important cinq jours plus tard.

Le flux médical dépend du jour de la semaine. Ce phénomène hebdomadaire se retrouve dans la sélection de la variable F_{t-7} . Le flux médical augmente le week-end et lundi, puis il diminue du mardi au vendredi. Cette augmentation est interprétable par le fait que les gens durant le week-end exercent des activités à risques : sport, jardinage, sorties... Suite à un accident, une grande partie de ces personnes se présentent directement aux urgences, car les médecins libéraux ne sont pas tous disponibles.

Pour ce qui est de l'influence du vent ou de l'humidité, les graphiques (d) et (e) montrent une très faible influence directe de ces deux variables explicatives sur le flux médical mais ces variables sont évidemment très corrélées avec la température. Notons cependant que l'humidité peut avoir des conséquences pour certaines pathologies pneumologiques ou rhumatologiques en particulier. Or la proportion de ces pathologies est très faible par rapport à la classe pathologie médicale où elles sont incluses. Leur variation n'aura donc que très peu d'influence sur le flux médical global.

Les données ainsi exposées, quelle relation peut-il y avoir avec la réalité médicale? Tout d'abord quels sont les problèmes que peuvent entraîner les températures élevées sur la santé? La déshydratation qui est une des conséquences des périodes chaudes, aura une expression clinique chez les individus les plus fragiles, en particulier les personnes âgées dont l'état général est précaire, les enfants en bas âge et les patients présentant déjà une tare. Cette déshydratation peut aussi décompenser de nombreuses autres pathologies telles les insuffisances cardiaques et les insuffisances rénales chroniques.

Certains troubles se déclenchent plus fréquemment quand il fait plus chaud, comme les infections et lithiases urinaires. Nîmes est une ville très ensoleillée. L'été y connaît donc une recrudescence de soucis dermatologiques à type de lucites (ou urticaire solaire), coups de soleil et autres tracas infectieux. Cette période correspond aussi à l'époque des vacances scolaires; période où les touristes sont très nombreux dans cette région. Pour des problèmes de médecine générale ils iront consulter plus facilement aux urgences qu'un médecin en ville.

L'apparition des pathologies que nous venons de citer, décompensations dues à la déshydratation ou dermatologiques, est souvent rapide après le facteur climatique déclenchant pris en compte. En fait la latence entre ce facteur et la consultation aux urgences est due essentiellement à la prise de décision tardive pour consulter, voir le graphique (c).

L'hiver, par contre, plus les températures sont basses et plus l'influence est significative sur les admissions des pathologies médicales dans le sens de l'augmentation au jour $J + 5$, voir le graphique (b). Cette influence n'est pas directe. L'hiver est la saison où les maladies infectieuses sont plus importantes. En effet, c'est la période des infections de la sphère Oto-Rhino-Laryngée (ORL) et pulmonaires, d'origine virale ou bactérienne. Le froid en est souvent le facteur déclenchant.

Cependant ce ne sont pas ces infections qui sont les motifs d'admissions mais plutôt leurs conséquences sur les personnes là aussi les plus fragiles; ceux dont l'état général est altéré par l'âge ou par une pneumopathie chronique déjà existant. Ici la latence est plus grande entre le facteur déclenchant et le jour de l'admission aux urgences. Ceci peut se comprendre par l'évolution en crescendo des maladies infectieuses sur plusieurs jours puis la dégradation clinique du patient entraînant son passage aux urgences.

De ces facteurs, seule la température semble avoir une influence significative. Nous avons vu que lors des périodes chaudes, le flux médical augmente surtout deux jours après une élévation de la température et cinq jours plus tard l'hiver quand la température baisse.

Les résultats étaient en partie prévisibles. En effet, les périodes très chaudes et les périodes très froides entraînent obligatoirement une fragilité du corps humain, d'où une recrudescence des problèmes médicaux. Par contre, la latence chiffrée entre le facteur déclenchant pris en compte et l'admission aux urgences est un renseignement nouveau.

Il faudra cependant pour une meilleure interprétation des courbes, avoir ce flux sur plusieurs années et les confirmer par une superposition.

7. Conclusion

L'ACPVI spline additive est une méthode performante dans l'étude des relations entre deux groupes de variables. Elle permet, même lorsque les variables explicatives sont très corrélées, de déterminer les variables influant sur les réponses. L'application de cette méthode à l'étude des séries chronologiques n'a, à notre connaissance, jamais été proposée. Les résultats des simulations permettent de préconiser cette méthode non paramétrique même lorsque le nombre des données est faible ou lorsque le nombre des variables explicatives est élevé.

Nous avons vu également l'intérêt de cette méthode pour l'identification des modèles. Ainsi outre que l'ACPVI spline additive est très performante pour un certain nombre de données (tout dépend évidemment des variables explicatives), la reconstruction des fonctions coordonnées, même pour n petit, permet de donner plusieurs pistes.

Pour terminer, notons que cette méthode est également applicable aux séries temporelles multivariées.

Remerciements

Nous remercions P.A. Cornillon pour toutes les discussions passionnées que nous avons eues ensemble. Nous remercions également les membres du comité de rédaction et les lecteurs pour leurs commentaires détaillés et constructifs.

8. Références

- AUESTAD B. et TJØSTHEIM D. (1990), Identification of nonlinear time series : first order characterization and order determination, *Biometrika*, **77**, 4, 669–687.
- BOSQ D. (1979), Sur la prédiction non paramétrique de variables aléatoires et mesures aléatoires, *Pub. interne, U.E.R de Mathématiques, Lille*.
- BREIMAN L. et FREIDMAN J.F. (1985), Estimating optimal transformations for multiple regression and correlation, (with discussion), *Journal of the American Statistical Association*, **80**, 580–618.
- CHEN R. et TSAY R.S. (1993), Nonlinear additive ARX models, *Journal of the American Statistical Association*, **88**, 298–308.
- CHENG B. et TONG H. (1992), On consistent nonparametric order determination and chaos, *Journal of the Royal Statistical Society, Series B*, **54**, 427–449.
- CIUPERCA G. (1997), Prédiction de l'ozone en région Parisienne, *XXIX^e Journées de Statistique ASU*, 265–266.
- COLLOMB G. (1980), Prédiction non paramétrique : étude de l'erreur quadratique du predictogramme, *Pub. interne, Univ. P. Sabatier, Toulouse*.
- COLLOMB G., HÄRDLE W. et HASSANI S. (1987), A note on prediction via estimation of the conditional mode function, *Journal of Statistical Planning and Inference*, **15**, 227–236.
- DURAND J.F. (1993), Generalized principal component analysis with respect to instrumental variables via univariate spline transformations, *Computational Statistics and Data Analysis*, **16**, 423–440.
- ESCOUFIER Y. (1987), Principal components analysis with respect to instrumental variables, *European Courses in Advanced Statistics, University of Napoli*, 285–299.
- GANNOUN A. (1989), *Estimation de la médiane conditionnelle*. Thèse de Doctorat de l'Université de Paris VI.
- HASTIE T. et TIBSHIRANI R. (1990), *Generalized additive models*, London, Chapman and Hall.
- IMAM W. et DURAND J.F. (1997), Une extension spline additive de l'Analyse en Composantes Principales sur Variables Instrumentales, *XXIX^e Journées de Statistique ASU*, 468–471.
- JONES D.A. (1978), Non-linear autoregressive processes, *Journal of the Royal Statistical Society, Series A*, **360**, 71–95.
- MASRY E. et TJØSTHEIM D. (1995), nonparametric estimation and identification of nonlinear time series, *Econometric Theory*, **11**, 258–289.
- RAO C.R. (1964), The use and the interpretation of principal component analysis in applied research, *Sankhya A*, **26**, 329–356.
- ROBINSON P.M. (1983), Nonparametric estimators for time series, *Journal of Time Series Analysis*, **4**, 185–207.
- SCHUMAKER L.L. (1981) *Spline Functions : Basic Theory*, Wiley, New York.

- TIBSHIRANI R. (1988), Estimating transformations for regression via additivity and variance stabilization, *Journal of the American Statistical Association*, **83**, 394–405.
- TJØSTHEIM D. et AUESTAD B. (1994a), Nonparametric identification of nonlinear time series : projections, *Journal of the American Statistical Association*, **89**, 1398–1409.
- TJØSTHEIM D. et AUESTAD B. (1994b), Nonparametric identification of nonlinear time series : selecting significant lags, *Journal of the American Statistical Association*, **89**, 1410–1419.
- TONG H. (1995), *Non linear Time Series*, London, Oxford.
- VIEU P. (1995), Order choice in nonlinear autoregressive models, *Statistics*, **26**, 307–328.