REVUE DE STATISTIQUE APPLIQUÉE

L. LÉGER
J. J. DAUDIN

Étude d'un modèle de régression non paramétrique : la régression par directions révélatrices

Revue de statistique appliquée, tome 41, n° 3 (1993), p. 21-48
<http://www.numdam.org/item?id=RSA_1993__41_3_21_0>

Rev. Statistique Appliquée, 1993, **XLI** (3), 21-48

ÉTUDE D'UN MODÈLE DE RÉGRESSION NON PARAMÉTRIQUE : LA RÉGRESSION PAR DIRECTIONS RÉVÉLATRICES

L. Léger (1), J.J. Daudin (2)

(1) (ENITAC, Chaire de Statistique-Informatique)
(2) (INA-PG, Chaire de Mathématiques-Informatique)

RÉSUMÉ

Les méthodes de régression non paramétrique se sont considérablement développées depuis 1980 : Régression par directions révélatrices (Friedman, 1981, 1984 a et b), modèles additifs généralisés (Hastie, Tibshirani, 1986, 1987, 1990), méthode A.C.E. (Breiman, Friedman, 1985). Il n'est pas toujours facile de savoir ce que recouvrent ces méthodes, ce qu'elles peuvent apporter et quelles en sont les limites. Notre objectif est de présenter ces techniques, d'illustrer leur fonctionnement sur des exemples et de proposer des critères de choix de modèle, de façon à permettre à un utilisateur n'ayant aucune expérience dans ce domaine de savoir ce qu'il peut en attendre. Il peut également permettre d'accéder de façon ordonnée à une bibliographie abondante et touffue.

Mots-clés : *Régression par Directions révélatrices, Modèles additifs, Méthodes de rééchantillonnage, Sélection de modèles-Erreur de prédiction.*

SUMMARY

Non parametric regression methods have been considerably developed since 1980 : Projection Pursuit Regression (PPR) (Friedman, 1981, 1984 a, b), Generalized Additive Models (GAM) (Hastie, Tibshirani, 1986, 1987, 1990), Alternating Conditional Expectation (ACE) (Breiman, Friedman, 1985). It's not always easy to understand their possibilities and their limits. In this paper, our objective is to present these techniques, to illustrate them on some worked examples and simulations and to propose some model selection criteria, which can be useful for an inexperienced user.

Key-words : *Projection Pursuit Regression, Additive models, Resampling methods, Model selection-Prediction error.*

1. Introduction

La régression non paramétrique (Eubank, 1988) a récemment été développée au cas multidimensionnel (Hastie, Tibshirani, 1990), c'est-à-dire à l'étude d'une

fonction de régression de $\mathbb{R}^p$, $f(x) = E[Y|X = x]$ p-dimensionnelle (Y désigne la variable réponse et $X = (X_1, .., X_p)$ un vecteur de variables explicatives) et propose actuellement toute une gamme de modèles possibles. Nous étudions dans cet article l'une de ces modélisations, appelée «Régression par Directions Révélatrices», traduction française du terme «Projection Pursuit Regression» (P.P.R.), (Friedman, Stuetzle, 1981).

L'approche non paramétrique considère le problème de l'estimation de f, sur la seule base d'hypothèses d'existence et de régularité pour f, et se caractérise par une grande souplesse de modélisation. Elle est à privilégier en particulier dans les situations où l'on ne dispose que de peu d'informations sur le type de liaison entre Y et X donc sur le modèle mis en jeu. Si la plus grande flexibilité des méthodes non paramétriques permet d'élargir le champ des situations modélisables, elles ne sont pas exemptes de difficultés (Collomb, 1981). L'une de ces difficultés concerne l'extension des techniques non paramétriques, initialement introduites dans le cas unidimensionnel ($p = 1$), au cas multidimensionnel, difficulté qui est à la base du développement de modèles tels que PPR. Dans la partie 2 nous explicitons brièvement ce problème ainsi que les différents modèles qui tentent d'y apporter une solution, modèles regroupés sous le nom de «régression non paramétrique additive» (Stone, 1985) La partie 3 présente les différentes versions des modèles PPR (Friedman, Stuetzle, 1981, Friedman, 1984 b) et leurs caractéristiques. Dans les parties 4 et 5 nous considérons l'étude d'un tel modèle par des méthodes de rééchantillonnage telles que le «bootstrap» (Efron, 1979) : nous traitons le problème du choix de modèles et abordons celui de l'inférence. La sixième partie présente quelques simulations suivies d'une discussion (partie 7).

2. La régression non paramétrique additive

2.1 Le problème posé par le cas multidimensionnel

L'extension des méthodes non paramétriques de régression telles que le lissage par fonction splines (Silverman, 1985) ou la régression par la méthode du noyau (Collomb, 1981), au cas multidimensionnel, c'est-à-dire l'estimation d'une fonction de régression f de $\mathbb{R}^p$ dans $\mathbb{R}$, sur la base d'un n-échantillon $(x_k, y_k), 1 \leqslant k \leqslant n$, pose un problème que les auteurs anglo-saxons désignent sous le nom de «curse of dimensionality» (Huber, 1985). Nous préciserons ce point, en envisageant le cas de la méthode à noyau. Cette méthode propose, dans le cas unidimensionnel un estimateur obtenu sous forme de moyennes locales pondérées des observations y_k, $1 \leqslant k \leqslant n$, correspondantes aux x_k situées dans un voisinage de x, voisinage défini par une fenêtre b_n, avec des poids fonction de la proximité de x_k à x et décroissants avec l'éloignement.

Dans le cas multidimensionnel, si la dimension p de l'espace des observations augmente, la taille de l'échantillon restant fixée, il est alors nécessaire pour maintenir une variance des estimations acceptable de moyenner sur un très large voisinage d'observations (boule de centre x et de rayon b_n), ce qui entraîne un biais croissant. Inversément, on montre (Eubank, 1988) que la variance des estimations est proportionnelle à $1/(n \times \lambda^p)$, où λ ($\lambda < 1$) représente un paramètre de lissage : si p croît, λ^p

tend vers 0, la variabilité croît alors rapidement, à moins de disposer d'échantillons de dimension gigantesque.

Ces difficultés sont essentiellement dues au fait qu'un espace de dimension élevé est «vide» (Huber, 1985), c'est-à-dire de très faible densité. Considérons par exemple n points distribués selon la loi uniforme, définie sur l'hypercube unité de $\mathbb{R}^p$ et considérons un cube contenant en moyenne 5% des points. Si $p = 2$, le coté de ce cube est 0.22, si $p = 50$ le coté est 0.94. Il s'en suit ce phénomène du «curse of dimensionality», lié à la dégradation des performances, lorsque on essaie d'étendre aux distributions de $\mathbb{R}^p$, avec p élevé, (en pratique dès que p dépasse quelques unités) les méthodes non paramétriques précédentes. Ce problème est à l'origine du développement des différents modèles de la régression non paramétrique additive, présentés ci-après.

2.2 Le modèle non paramétrique additif

La régression dite non paramétrique additive (Stone, 1985) regroupe une gamme de modèles qui proposent, dans les cas où l'on ne dispose pas d'échantillons de taille suffisante pour envisager une approche p-dimensionnelle directe, une approximation additive pour f, du type :

$$f(x_1, \ldots, x_p) = \alpha + \sum_{j=1}^{p} f_j(x_j)$$

Les fonctions $f_j, 1 \leqslant j \leqslant p$, sont des fonctions réelles d'une seule variable réelle, de type fonctions lisses définies point par point et non de forme analytique fixée *a priori*. Ce modèle constitue une extension naturelle de différents modèles paramétriques, prenant en compte trois aspects fondamentaux d'un modèle de régression (Stone, 1985) : aspects liés à la double qualité requise pour le modèle, explicative et prédictive.

Le premier est la flexibilité du modèle, ou capacité de modéliser des situations de nature différente. L'introduction de fonctions lisses permet d'envisager, en particulier, l'étude de situations de dépendance non linéaire. L'absence de flexibilité est à relier en ce sens à un risque possible de biais du modèle estimé, en particulier dans les cas où l'on ne dispose que de peu d'informations sur la fonction de régression. Une telle flexibilité doit cependant avoir ses limites : en effet si on n'impose aucune contrainte à l'approximation, il est possible de faire passer une courbe estimée par tous les points de l'échantillon observé, avec une variance résiduelle nulle. Mais alors dans ce cas, l'estimation obtenue n'aura aucune valeur prédictive. Le second aspect est la «dimension» du modèle, lié à la variance des estimations, qui croit rapidement pour n fixé, si p augmente (problème de la «curse of dimensionality») et entraîne donc l'instabilité du modèle estimé. La structure additive du modèle, avec estimation de fonctions univariées, propose une solution à ce problème. Les deux aspects, flexibilité et «dimension» sont étroitement liés et la recherche d'un modèle correct peut se baser, comme nous le verrons au paragraphe 5, sur un compromis contrebalançant flexibilité et dimension. Le troisième aspect est l'interprétabilité du modèle, importante pour la compréhension de la structure probabiliste sous jacente. La représentation des

fonctions f_j, $1 \leqslant j \leqslant p$, favorise une telle interprétation en révélant la relation entre Y et un régresseur particulier, conditionnellement à la présence des autres régresseurs.

Différentes extensions de ce modèle de base ont été développées, toutes basées sur une hypothèse d'additivité de l'approximation, plus précisément :

– Les modèles Additifs Généralisés (GAM) (Hastie, Tibshirani, 1986 et 1987) :

$$g(E[Y|X = x]) = g(f(x_1, \ldots, x_p)) = \sum_{j=1}^{p} f_j(x_j)$$

qui constituent une extension des modèles linéaires généralisés, avec étude de f au travers d'une fonction lien g.

– Les modèles de «Régression par Directions Révélatrices» (PPR) :

$$E[Y|X = x] = f(x_1, \ldots, x_p) = \sum_{m=1}^{M} f_m(\sum_{j=1}^{p} b_{jm} x_j)$$

qui font l'objet de cet article.

– Méthode «Alternating Conditional Expectation» (ACE) (Breiman et Friedman, 1985) :

$$E[\phi(Y)|X = x] = \sum_{j=1}^{p} f_j(x_j)$$

qui proposent une double transformation, de la variable réponse et des régresseurs.

3. Régression par directions révélatrices

3.1 Les méthodes de directions révélatrices en régression

Les méthodes de «Directions Révélatrices ont tout d'abord été introduites en analyse exploratoire avec l'expérience PRIM9 de Friedman et Tuckey (1974). Elles ont ensuite été développées dans le domaine de l'estimation fonctionnelle multivariée, estimation d'une fonction de régression avec la méthode de «Projection Pursuit Regression» (Friedman et Stuetzle, 1981) et estimation d'une densité avec la méthode de «Projection Pursuit Density Estimation» (Friedman, Stuetzle, Schroeder, 1984). Ces différentes extensions ont été formalisées par Huber (1985) qui en a dégagé les idées fortes et proposé un cadre commun. Dans le domaine de l'estimation non paramétrique, leur principal intérêt est de proposer une solution au problème de la «curse of dimensionality».

Nous nous intéressons dans cet article au cas de la régression, c'est-à-dire aux modèles PPR, basés sur une approximation de la fonction de régression du type, pour Y centrée :

$$f(x_1, x_2, \ldots, x_p) = \sum_{m=1}^{M} f_m(\alpha'_m x) \text{ avec } \alpha'_m x = \sum_{j=1}^{p} \alpha_{jm} x_j.$$

Les fonctions $f_m (1 \leqslant m \leqslant M)$ sont des fonctions réelles d'une variable réelle, combinaison linéaire des variables $X_j, 1 \leqslant j \leqslant p$, définie par le vecteur unitaire ou direction α_m, $1 \leqslant m \leqslant M$. L'idée est donc d'étudier une fonction de p variables au travers de son approximation par p fonctions d'une seule variable.

Ces modèles constituent une application des Directions Révélatrices à la régression. L'approximation de la fonction de régression est basée sur la projection des observations sur des plans engendrés par la variable Y et une combinaison linéaire $\alpha' X$ des variables explicatives. La flexibilité de l'approximation et la substitution des variables initiales par des combinaisons $Z_m = \alpha'_m X$ permettent l'étude de situations de dépendance non linéaire, avec d'éventuelles interactions entre régresseurs. Un autre avantage est celui de l'équivalence affine de la solution déterminée par Directions Révélatrices : invariance par toute transformation affine non singulière, rotation et changement d'échelle sur les régresseurs initiaux. L'estimation de fonctions univariées $f_m, 1 \leqslant m \leqslant M$, permet d'éviter les difficultés de la «curse of dimensionality». Diaconis et Shashahani (1984) ont montré en outre que toute fonction de p variables peut être approchée par un développement de type PPR, pour M assez grand.

Une deuxième version des modèles PPR a été proposée par Friedman (1984 a), baptisée «Smooth Multiple Additive Regression Technique» (en abrégé S.M.A.R.T.) qui permet de modéliser un ensemble de q variables réponses Y_i, $1 \leqslant i \leqslant q$, en fonction de p variables prédictrices X_j, $1 \leqslant j \leqslant p$, sous la forme :

$$E[Y_i | x_1, \ldots, x_p] = \beta_{i0} + \sum_{m=1}^{M} \beta_{im} f_m(\alpha'_m x) \text{ et } \alpha'_m x = \sum_{j=1}^{p} \alpha_{jm} x_j$$

avec les contraintes suivantes, pour $1 \leqslant m \leqslant M$:

$$E[f_m(\alpha'_m X)] = 0, \; E[f_m^2(\alpha'_m X)] = 1 \text{ avec } \alpha_m \text{ unitaires } (\sum_{j=1}^{p} \alpha_{jm}^2 = 1).$$

Chaque variable réponse est modélisée sous la forme d'une combinaison linéaire (coefficients β_{im}) de fonctions prédictrices $f_m (1 \leqslant m \leqslant M)$.

Il existe également une troisième variante baptisée Multidimensional Adaptative Splines Approximation (en abrégé M.A.S.A.) (Friedman, Grosse, Stuetzle, 1983), utilisant un lissage par fonctions splines et une recherche des directions révélatrices par un algorithme de Gauss.

3.2 Méthode d'estimation

Nous décrivons ici brièvement le principe de la méthode d'estimation des paramètres, directions et fonctions , sur la base d'un n-échantillon $(y_k, x_k), 1 \leqslant k \leqslant n$.

Etape 1 : on détermine un premier couple $(\widehat{\alpha}_1, \widehat{f}_1)$, par la minimisation d'un critère des moindres carrés :

$$L_2 = \sum_{k=1}^{n} [y_k - f_1(\alpha_1' x_k)]^2.$$

On substitue ensuite aux y_k les résidus courants définis par :

$$r_k = y_k - \widehat{f}_1(\widehat{\alpha}_1' x_k) \text{ pour } 1 \leqslant k \leqslant n.$$

et on recherche un second couple $(\widehat{\alpha}_2, \widehat{f}_2)$ par minimisation du même critère. On itére ainsi le processus : nous décrivons ici l'étape m.

Etape m : on pose $r_k = y_k - \sum_{l=1}^{m-1} f_l(\widehat{\alpha}_l' x_l)$ $1 \leqslant k \leqslant n$, et on détermine $(\widehat{\alpha}_m, \widehat{f}_m)$ par minimisation de :

$$L_2 = \sum_{k=1}^{n} [r_k - f_m(\alpha_m' x_k)]^2.$$

La fonction f_m est obtenue, pour une direction fixée α_m sous la forme :

$$f_m(z) = E[r(X) | \alpha_m' X = z].$$

Le critère L_2 est ensuite minimisé pour tous les choix possibles de la direction α_m de $\mathbb{R}^p$, exploration réalisée grâce à un algorithme de Rosenbrock restreint à la sphère unité de $\mathbb{R}^p$.

Le processus est ainsi réitéré jusqu'à convergence, c'est-à-dire jusqu'à ce qu'ajouter un terme supplémentaire n'améliore pas significativement le critère, ou encore jusqu'à ce qu'il n'y ait plus de structure dans les résidus. Il n'existe toutefois pas de garantie de convergence vers le minimum global de L_2.

Dans la seconde version (se reporter à Friedman (1984 a) pour le détail) il est en outre prévu un réajustement des termes par plusieurs passages de l'algorithme, procédure dite de «backfitting» (analogue à la procédure «stepwise» en régression ascendante). Un terme quelconque de la décomposition peut alors être réestimé aprés détermination des autres termes. On pourra se reporter à Buja, Hastie et Tibshirani (1989) pour une discussion de l'amélioration due au «backfitting».

3.3 Procédure de lissage

L'estimation des fonctions f est obtenue par une procédure de lissage de type méthode à noyau. Les fonctions sont déterminées point par point, par ajustement d'une courbe de lissage aux résidus courants r_k ou $y_k, 1 \leqslant k \leqslant n$, sur la direction correspondante. Le problème essentiel est alors le choix de la fenêtre de lissage. Si elle est trop petite, on obtiendra une estimation irrégulière de variance élevée; si elle est trop grande, le biais sera élevé, l'estimation étant trop lissée. En outre étant donné la structure particulière des modèles PPR, un surajustement sur l'une des premières directions, peut invalider les directions suivantes. L'idée commune aux méthodes de détermination de la fenêtre est de contrebalançer idéalement les effets du biais et de la variance.

Dans la première version, Friedman et Stuetzle ont choisi un ajustement linéaire local, permettant de réduire le biais éventuel d'une simple procédure par moyennes mobiles constantes. Une estimation pour des valeurs z différentes de l'une des valeurs z_k est obtenue par simple interpolation linéaire. La fenêtre est variable sur chacune des directions estimées. On choisit une valeur moyenne, qui peut ensuite évoluer en fonction de la variabilité locale de la réponse, avec une fenêtre plus grande lorsque la variabilité est plus élevée.

La seconde version est elle basée sur une procédure de lissage baptisée «supersmoother» (Friedman, 1984 a) et caractérisée par une détermination automatique, à partir des données, de la fenêtre et ce indépendamment sur chacune des directions estimées. Cette détermination est basée sur la minimisation d'une erreur quadratique de modélisation $E[Y - f(X)]^2$ ou $E[E[Y|X] - f(X)]^2$, obtenue par une méthode de validation croisée (Stone, 1974), dont nous reparlerons au paragraphe 4. Une détermination automatique du paramètre de lissage améliore sensiblement l'estimation mais rend impossible le traitement analytique du modèle estimé, la procédure devenant alors non linéaire (Buja, Hastie, Tibshirani, 1989).

4. Etude des modèles : critères de sélection

La régression par Directions Révélatrices propose une hiérarchie de modèles de complexité croissante, liée au nombre de termes M mis en jeu. On a vu que toute fonction de p variables peut être approchée par un développement du type PPR pour M suffisamment grand. Toutefois inclure trop de termes dans un tel développement entraîne des difficultés de stabilité et d'interprétabilité pour le modèle estimé.

Le problème posé est donc celui de la sélection d'un modèle ou choix de M. Nous considérons dans la suite le cas de la version Friedman (1984 a) avec une seule variable réponse $Y(q = 1)$. L'approche que nous proposons (Daudin, Léger, 1989) est inspirée de celle introduite par Linhart et Zucchini, (1985) dans un contexte plus général incluant l'analyse de régression. Cette approche est basée sur la notion de dissemblance entre modèle théorique et modèle estimé.

4.1 Procédure de sélection

La dissemblance que nous choisirons pour la sélection de modèles est celle dite de prédiction, qui fournit un bon jugement sur les performances d'un modèle. Certains critères de sélection introduits en régression linéaire multiple comme la statistique de Mallows ou le critère Press (Hocking, 1976) peuvent être considérés comme construits à partir d'une telle dissemblance.

Nous noterons G_θ avec $\theta \subset \mathbb{R}^p$ la famille d'approximation des modèles PPR, F le modèle opératoire, ou plus proche représentation probabiliste de la situation qu'il est possible d'envisager (en l'absence de toute information *a priori* sur f, ce modèle n'est précisé que par le seul échantillon observé).

La dissemblance de prédiction se définit alors comme l'erreur quadratique de prédiction s'écrivant, dans le cas où X est un vecteur aléatoire de $\mathbb{R}^p$:

$$\Delta(G_\theta, F) = E_F[Y - E_{G_\theta}[Y|x]]^2.$$

Cette dissemblance se décompose formellement comme la somme de deux termes : un premier terme lié au biais du modèle (dissemblance dite d'approximation) et un terme lié a la variance des estimations (ou dissemblance d'estimation). Le terme de biais est spécifique du modèle choisi et donc en particulier du nombre de termes M. Son effet diminue rapidement lorsque M augmente. Le terme lié à la variance des estimations au contraire augmente régulièrement lorsque M croît. Intuitivement il existe donc une valeur optimale de M pour un ensemble de données connues (avec n fixé).

Il est donc possible de développer une sélection de modèles basée sur la minimisation d'un critère, qui se définit comme tout estimateur de $E_F[\Delta(G_{\hat{\theta}}, F)]$; ce qui revient à choisir la procédure qui en moyenne conduit à la plus faible dissemblance.

La construction d'un critère requiert donc une estimation de l'expression précédente. Bunke et Droge (1984 a et b) et Droge (1987) ont déja utilisé une telle approche en régression linéaire et en régression non linéaire. Dans une situation non standard, comme celles des modèles PPR, en l'absence d'une expression explicite de $E_F[\Delta(G_{\hat{\theta}}, F)]$, se pose le problème de la détermination d'une telle estimation. Linhart et Zucchini proposent en ce cas d'utiliser des techniques de rééchantillonnage qui permettent d'estimer cette espérance. Ces techniques présentent en outre l'avantage de pouvoir être appliquées pour un large choix de dissemblances.

4.2 Un premier critère de sélection

Un premier estimateur intuitif se base sur la somme des carrés résiduels correspondant à l'échantillon observé (y_k, x_k), $1 \leqslant k \leqslant n$, et s'écrit sous la forme :

$$\widehat{r}_{AE} = || y - \widehat{y} ||_V^2 = \sum_{k=1}^{k=n} v_k (y_k - \widehat{y}_k)^2.$$

où $\widehat{y}$ représente le vecteur des estimations obtenues et $V = (v_k)$, $1 \leqslant k \leqslant n$, la matrice diagonale des poids des observations (on prendra dans la suite des poids tous égaux à $1/n$).

Le choix de modèles consiste alors en une inspection des valeurs du critère pour différentes valeurs de M (équivalent à la recherche du minimum de L_2). Toutefois à cause de l'existence de possibles minimums locaux, pour une valeur de M donnée, les solutions du problème posé ne sont pas nécessairement obtenues dans l'ordre M croissant. Tomassone, Danzart, Daudin, Masson (1988) proposent alors un choix de M correspondant aux sauts brusques du critère par sélection descendante. On part d'une valeur $M = M_{\text{Sup}}$ et on examine successivement $M_{\text{Sup}}, M_{\text{Sup}-1}, \ldots, M_{\text{Inf}}(=?1)$. L'importance de chaque terme est mesurée par l'expression : $I_m = |\beta_m|, 1 \leqslant m \leqslant M$ (dans le cas $q = 1$, on posera $\beta_m = \beta_{im}$ dans l'expression de $E[Y_i|x_1, \ldots, x_p]$ figurant page 8) importance normalisée, le terme le plus important étant d'importance 1.

Ce critère correspond à une estimation de l'erreur de prédiction dite apparente, car déterminée sur les mêmes observations $(y_1, \ldots, y_n)$, ayant permis l'estimation du modèle. Il présente toutefois un inconvénient majeur; il est biaisé négativement et sous-estime la véritable erreur de prédiction.

L'idée est alors la construction d'un critère corrigé du biais. Bunke et Droge l'ont proposé, dans le cas de la régression linéaire, à partir d'une décomposition explicite du critère. Toutefois dans le cas PPR, en l'absence d'une telle décomposition, nous développerons une variante introduite par Efron (1983) en discrimination, basée sur des techniques de rééchantillonnage.

4.3 Critères bootstrap et validation croisée

Les critères développés sont basés sur une amélioration du biais du critère $\widehat{r}_{AE}$. Le biais de l'erreur apparente, noté $R(\widehat{F}, F)$, se définit sous la forme :

$$R(\widehat{F}, F) = E_F[\Delta(z, g_{\hat{\theta}}(x_z))] - E_{\hat{F}}[\Delta(z, g_{\hat{\theta}}(x_z))]$$

où $\Delta(z, g_{\hat{\theta}}(x_z))$ représente la dissemblance de prédiction causée par la prédiction $g_{\hat{\theta}}(x_z) = E_{G\widehat{\theta}}[Z|x_z]$ d'une nouvelle valeur (Z, X_z) indépendante des (Y_k, X_k), $1 \leqslant k \leqslant n$. Le premier terme de l'expression représente l'espérance de la dissemblance de prédiction :

$$E_F[\Delta(z, g_{\hat{\theta}}(x_z))] = E[Z - E_{G_{\hat{\theta}}}[Z|x_z]]^2.$$

Le deuxième terme de l'expression correspond à l'erreur apparente sur l'échantillon (x_k, y_k), $1 \leqslant k \leqslant n$, et il s'écrit :

$$E_{\hat{F}}[\Delta(z, g_{\hat{\theta}}(x_z))] = \frac{1}{n} \sum_{k=1}^{k=n} \Delta(y_k, g_{\hat{\theta}}(x_k)).$$

La procédure de rééchantillonnage s'applique ici à l'étude de la distribution de la v.a. $R(\widehat{F}, F)$ pour obtenir une estimation de son espérance (espérance du biais de l'erreur) :

$$\omega = E_{\hat{F}\sim F} R(\widehat{F}, F)$$

où l'expression $E_{\hat{F}\sim F}$ désigne l'espérance prise par rapport à $\widehat{F}$, obtenue à partir d'un échantillon généré de loi F.

a) Critère bootstrap

La méthode du bootstrap (Efron, 1979, 1983a et b, 1988) consiste en un rééchantillonnage des observations préservant la structure probabiliste initiale, pour construire de pseudos-données, ou échantillon «bootstrap» sur lesquelles on étudiera la statistique considérée. L'application de la méthode à l'analyse de régression a été développée par Freedman (1981). Deux modes de rééchantillonnage sont possibles, un premier consistant en un rééchantillonnage des résidus et le second, que nous utiliserons ici, consistant a rééchantillonner les vecteurs (Y, X) eux-mêmes. Le rééchantillonnage va permettre d'obtenir, à partir d'un échantillon bootstrap (Y_k^*, X_k^*), $1 \leqslant k \leqslant n$, de loi $\widehat{F}$ en remplaçant F par $\widehat{F}$ et $\widehat{F}$ par $\widehat{F}^*$ (où $\widehat{F}^*$ désigne la distribution empirique de l'échantillon bootstrap de distribution $\widehat{F}$), une estimation :

$$\widehat{\omega}_{\text{Boot}} = E_{\hat{F}^*\sim\widehat{F}} R(\widehat{F}^*, \widehat{F}).$$

Cette expression peut être approchée par un algorithme de Monte-Carlo sous la forme suivante :

– Etape 1 : On génère un échantillon bootstrap $(Y_k^*, X_k^*), 1 \leqslant k \leqslant n$ de distribution $\widehat{F}$.

– Etape 2 : On calcule à partir de l'estimation de la fonction de régression $g_{\hat{\theta}^*}$ déterminée sur l'échantillon $(y_1^*, \ldots, y_n^*)$ le biais de l'erreur apparente $R^*(\widehat{F}^*, \widehat{F})$:

$$\begin{aligned} R^*(\widehat{F}^*, \widehat{F}) &= E_{\hat{F}}[\Delta(z, g_{\hat{\theta}}^*(x_z))] - E_{\hat{F}^*}[\Delta(z, g_{\hat{\theta}}^*(x_z))] \\ &= \frac{1}{n}\sum_k \Delta(y_k, g_{\hat{\theta}}^*(x_k)) - \frac{1}{n}\sum_k \Delta(y_k^*, g_{\hat{\theta}}^*(x_k^*)) \end{aligned}$$

différence entre erreur vraie et erreur apparente de prédiction sur l'échantillon bootstrap.

– Etape 3 : On répète les étapes 1 et 2 un nombre B de fois pour obtenir un B échantillon de la v.a. R :

$$R^{*1}, \ldots, R^{*B}.$$

L'approximation Monte-Carlo de $\widehat{\omega}_{\text{Boot}}$ s'écrit :

$$\widehat{\omega}_B = \frac{1}{B}\sum_{b=1}^{B} R^{*b}(\widehat{F}^*, \widehat{F}).$$

Le critère bootstrap se définit ensuite sous la forme suivante :

$$\widehat{r}_B = \widehat{r}_{AE} + \widehat{\omega}_B.$$

b) Critère Validation croisée

La méthode de validation-croisée (déjà mentionnée pour la détermination d'un paramètre de lissage), proposée par Stone (1974) permet également d'obtenir un estimateur du biais de l'erreur de prédiction. Son principe consiste à exclure une observation, à estimer la fonction de régression sur le $(n-1)$ échantillon restant et à déterminer la prédiction de l'observation exclue sur cette base. On notera $g_{\hat{\theta}(-k)}(.)$ l'estimation de la fonction de régression sur le $(n-1)$ échantillon (x_l, y_l), $1 \leqslant l \neq k \leqslant n$ dont l'observation (x_k, y_k) a été exclue et $\widehat{y}_{(-k)}$ la prédiction $g_{\hat{\theta}(-k)}(x_k)$ de y_k obtenue grâce à cette fonction de régression estimée. Ce processus est ensuite réitéré sur chaque observation. La moyenne des erreurs de prédiction ainsi obtenues constitue la mesure par validation croisée de l'erreur de prédiction (on prendra toujours des poids égaux à $1/n$). Le critère correspondant s'écrit :

$$\begin{aligned}\widehat{r}_{cv} &= ||y - \widehat{y}_{(-)}||_V^2 \text{ avec } \widehat{y}_{(-)} = (\widehat{y}_{(-1)}, \ldots, \widehat{y}_{(-n)}).\\ &= \frac{1}{n}\sum_{k=1}^{n} \Delta(y_k, g_{\hat{\theta}(-k)}(x_k))\end{aligned}$$

On déduit de ce critère une estimation de l'espérance du biais de l'erreur de prédiction :

$$\widehat{\omega}_{CV} = \frac{1}{n}\sum_{k=1}^{n} \Delta(y_k, g_{\hat{\theta}(-k)}(x_k)) - \frac{1}{n}\sum_{k=1}^{n} \Delta(y_k, g_{\hat{\theta}}(x_k))$$

ou différence entre les erreurs observées, selon que la $k^{\text{ième}}$ observation figure ou non dans l'échantillon.

5. Etude des modèles : traitement par rééchantillonnage

L'étude des propriétés des estimations obtenues, dont l'importance a été soulignée par Huber (1985) et Buja, Hastie, Tibshirani (1989), est pratiquement inexistante pour de tels modèles. Deux aspect peuvent être développés, le premier concerne l'étude de la stabilité et de l'interprétabilité des estimations, directions et fonctions. Le second est un problème de test sur la significativité de l'ajustement

proposé. Nous aborderons le premier point, nous limitant à mentionner quelques éléments bibliographiques pour le deuxième.

L'approche proposée est issue du rééchantillonnage bootstrap, (Efron, 1988) qui permet dans les cas où une analyse standard n'est pas possible ou conduit à des résultats, asymptotiques, d'obtenir certaines propriétés pour les estimateurs.

5.1 Etude des directions

L'étude des directions $(\widehat{\alpha}_m)$, $1 \leqslant m \leqslant M$, estimées peut être menée à partir des distributions bootstrap des estimateurs des contributions des différentes variables $X_1, \ldots, X_p$, comme le suggèrent Efron et Tibshirani (1986). Considérons $\alpha = (\alpha_j)$, $1 \leqslant j \leqslant p$, une direction quelconque, $\widehat{\alpha}$ son estimateur. L'espérance $E(\widehat{\alpha})$ (donc le biais) et la matrice de variance-covariance $V(\widehat{\alpha})$ peuvent être approchées par les caractéristiques correspondantes déterminées sur les histogrammes bootstrap des estimations $\widehat{\alpha}_j^{*(b)}$ $1 \leqslant j \leqslant p$, pour $1 \leqslant b \leqslant B$:

$$B^*(\widehat{\alpha}^*) = (1/B)\sum_{b=1}^{B}(\widehat{\alpha}^{*(b)} - \widehat{\alpha}) = \widehat{\alpha}^{*(.)} - \widehat{\alpha} \text{ (biais)}$$

$$V^*(\widehat{\alpha}^*) = (1/B-1)\sum_{b=1}^{B}(\widehat{\alpha}^{*(b)} - \widehat{\alpha}^{*(.)})(\widehat{\alpha}^{*(b)} - \widehat{\alpha}^{*(.)})^t \text{ (variance)}$$

L'examen de la matrice V^* permet de juger de la stabilité de la direction estimée, au travers des contributions.

5.2 Etude des fonctions

Le rééchantillonnage permet également la construction d'intervalles de confiances point par point, pour une fonction f_m particulière. La représentation graphique des f_m et des intervalles de confiance renseigne sur la variabilité de l'estimation et fournit une interprétation utile du modèle estimé.

Considérons une fonction f_m, d'estimation $\widehat{f}_m$ et notons $\widehat{f}_m^{*(b)}$, $1 \leqslant b \leqslant B$, les replications déduites du rééchantillonnage bootstrap, conditionnellement à une direction fixée $z_m = \alpha_m' x$. La construction d'intervalles de confiance nécessite une approximation de l'erreur quadratique $E(\widehat{f}_m(z_m) - f_m(z_m))^2$ ou l'espérance est prise sur plusieurs échantillons de taille n, issus de la même population. Une telle approximation est obtenue en approchant la distribution de l'expression $\widehat{f}_m(z_m) - f_m(z_m)$ par la distribution bootstrap de $\widehat{f}_m^{*(b)}(z_m) - \widehat{f}_m(z_m)$.

L'estimation de l'erreur quadratique précédente s'écrit dans ces conditions :

$$\widehat{E}(\widehat{f}_m(z_m) - f_m(z_m))^2 = E_B(\widehat{f}_m^{*(b)}(z_m) - \widehat{f}_m(z_m))^2$$

où E_B indique l'espérance prise par rapport aux B échantillons bootstrap.

Afin de tenir compte de l'asymétrie possible de la distribution de $\widehat{f}_m(z_m) - f_m(z_m)$, causée soit par l'asymétrie de la distribution des erreurs, soit par le biais de l'estimateur $\widehat{f}_m(z_m)$, Friedman et Silverman (1989) ont proposé d'utiliser la souplesse du bootstrap pour construire des intervalles de confiance asymétriques autour de la courbe. Pour ceci, on détermine à partir du faisceau de courbes estimées $\widehat{f}_m^{*(b)}(z_m)$, les deux expressions suivantes, pour chacune des valeurs de la variable Z sur l'échantillon initial :

$$e^2_{(1)}(z_m) = E_B^{(-)}(\widehat{f}_m^{*(b)}(z_m) - \widehat{f}_m(z_m))^2$$
$$e^2_{(2)}(z_m) = E_B^{(+)}(\widehat{f}_m^{*(b)}(z_m) - \widehat{f}_m(z_m))^2$$

correspondantes aux espérances prises par rapport aux replications pour lequelles, respectivement $\widehat{f}_m^{*(b)}(z_m) - \widehat{f}_m(z_m)$ est négatif (1) et $\widehat{f}_m^{*(b)}(z_m) - \widehat{f}_m(z_m)$ est positif (2). Les valeurs $e^2_{(1)}(z_m)$ et $e^2_{(2)}(z_m)$ obtenues pour chaque observation (ou valeur de Z) sont ensuite lissées en fonction de z_m. Un tel lissage peut être effectué ici par une procédure de type moyennes mobiles avec pas constant. Les courbes lissées, ainsi déterminées $\widehat{e}^2_{(1)}(z_m)$ et $\widehat{e}^2_{(2)}(z_m)$ sont utilisées pour définir les intervalles de confiance relatifs aux estimations $\widehat{f}_m(z_m)$ initiales :

$$[\widehat{f}_m(z_m) - \widehat{e}_{(1)}(z_m);\ \widehat{f}_m(z_m) + \widehat{e}_{(2)}(z_m)].$$

5.3 Discussion sur l'inférence

La difficulté de traitement d'un tel modèle (propriétés des estimateurs, tests sur la validité des termes de l'approximation) a conduit différents auteurs à introduire des modèles «idéalisés» capables de décrire la modélisation PPR mais sur lesquels pèsent certaines hypothèses qui en facilisent l'étude.

Hall (1989) a proposé un modèle mathématique permettant de démontrer l'équivalence entre l'estimateur à noyau PPR sur la direction estimée et un estimateur à noyau sur la direction, non aléatoire, déterminée analytiquement à partir de ce modèle. Il obtient ainsi des expressions explicites du biais et de la variance (complexes) des estimateurs des directions et des fonctions PPR. Ces expressions montrent que la variance de ces estimations est très voisine de celles d'une estimation univariée à noyau. Par contre les formules de biais sont beaucoup plus complexes, complexité due au biais dans l'estimation de la direction de projection.

Les difficultés précédentes sont également présentes dans le domaine des tests de significativité. Une procédure de rééchantillonnage pour la construction d'un test de type bootstrap ou de permutation, serait en outre très coûteuse en temps de calcul. Johansen et Johnstone (1990) ont, ici encore à partir d'un modèle idéalisé, proposé un test de significativité du premier terme $\widehat{f}_1(\widehat{\alpha}_1' x_k)$ du développement, qui peut formellement être généralisé aux termes suivants.

6. Simulations

Dans cette partie, nous présentons quelques simulations, effectuées avec la version PPR-SMART de Friedman (1984 a et b), permettant de juger de la capacité de PPR à modéliser des simulations de différente nature ainsi que des propriétés des estimations obtenues et de la qualité des critères. En ce qui concerne la procédure de lissage «supersmoother», nous avons constaté (Léger, 1991) sa très bonne adaptation aux situations très bruitées ou à courbure importante. Toutefois elle présente une tendance nette au surajustement, due à sa très (ou trop) grande adaptabilité.

6.1 Simulation 1

Ce premier exemple consiste en la simulation de $N = 50$ échantillons de taille $n = 100$, observations générées suivant le modèle :

$$Y = \exp 3(X_1 + X_2 - X_3) + 10X_4 + \varepsilon$$

avec X_1, X_2, X_3, X_4 v.a. de loi Uniforme $U[0,1]$ générées indépendamment et $\varepsilon \sim N(0; 0.01)$. Ces N échantillons vont nous permettre d'estimer l'espérance de la dissemblance de prédiction $E_F[\Delta(G_{\widehat{\theta}}, F)]$, base des diférents critères de sélection développés.

Les résultats obtenus pour cette sélection de modèles sont reportés au tableau 1.1 avec les valeurs des critères (E) et leur écart-type (σ). Le bootstrap du modèle est effectué avec $B = 50$ replications. Les modèles testés sont ceux à $M = 1, 2, 3$ et 4 termes. Le modèle retenu est celui à $M = 2$ termes. Nous avons également considéré un estimateur de l'erreur de prédiction, déterminé par prédiction sur la base de la fonction de régression estimée, d'un nouvel échantillon de $n' = 800$ observations. Un tel estimateur doit en effet constituer une estimation assez fidèle de l'erreur de prédiction.

Nous remarquons que l'erreur apparente $\widehat{r}_{AE}$ sous-estime largement l'erreur r (biais important). L'estimation du coefficient de corrélation multiple dans ce type de modèles est donc beaucoup trop optimiste. Les estimateurs $\widehat{r}_{CV}$ et $\widehat{r}_B$ améliorent nettement l'estimation $\widehat{r}_{AE}$: on note toutefois dans certains cas l'existence de minimums locaux qui font que le bon modèle n'est pas trouvé. La variabilité des critères est élevée. Cette variabilité de $\widehat{r}_{CV}$ et $\widehat{r}_B$ peut être décomposée sous forme de somme de deux termes : un premier terme de variance dû à l'erreur dans l'estimation de l'erreur de prédiction d'une règle de prédiction fixée et un deuxième terme dû à l'erreur entre différentes règles basées sur des échantillons différents de la même population. Il faut enfin noter une variabilité plus élevée de $\widehat{r}$. Ce phénomène peut être expliqué à partir de la notion de «distance» entre échantillons : en effet $\widehat{r}_{CV}$ et $\widehat{r}_B$ sont par construction déterminés à partir d'un rééchantillonnage de l'échantillon initial, alors que $\widehat{r}$ est elle déterminée sur un nouvel échantillon qui peut parfois être beaucoup plus éloigné de cet échantillon initial.

La deuxième partie de la simulation consiste en une étude, sur l'un des $N = 50$ échantillons, des propriétés des estimations obtenues, directions et fonctions, grâce au rééchantillonnage bootstrap du modèle.

TABLEAU 1.1
Critères $\widehat{r}_{AE}$, $\widehat{r}_{CV}$, $\widehat{r}_B$ et estimation $\widehat{r}$ de l'erreur de prédiction, espérance (E), écart-type(σ)

Critères		M=1	M=2	M=3	M=4
$\hat{r}_{AE}$	(E)	10.10	3.45	3.23	2.88
	(σ)	7.51	5.06	4.88	3.97
$\hat{r}_{CV}$	(E)	15.15	7.27	7.90	7.75
	(σ)	7.57	7.00	6.28	5.80
$\hat{r}_{B}$	(E)	17.05	7.76	8.09	8.26
	(σ)	11.90	7.10	6.81	7.63
$\hat{r}$	(E)	15.23	7.09	6.83	6.63
	(σ)	10.41	9.05	9.46	9.28

L'étude des distributions bootstrap ($B = 50$ replications) des estimateurs des directions est reportée sur les tableaux 1.2, 1.3 et sur les histogrammes 1.4. Le tableau 1.2 donne les estimations des contributions des X_j, $1 \leqslant j \leqslant p$ pour les deux premières directions, ainsi que les espérances et écarts-type de ces mêmes estimateurs obtenus par bootstrap. Les importances des variables X sont mesurées pour $q = 1$ par :

$$I_j = \sigma_j E|\sum_m \beta_m \alpha_{jm} f'_m(\alpha'_m X)|,\ 1 \leqslant j \leqslant p, \text{ (cf. Friedman, 1984 b)}$$

(où σ_j représente un facteur d'échelle, α_{jm} la $j^{\text{ième}}$ composante de α_m et f'_m la dérivée de f_m) : leurs valeurs sont ici : $1.0(X_1)$, $0.96(X_2)$, $0.94(X_3)$, $0.20(X_4)$. L'examen du tableau révèle la relative stabilité des contributions de chacun des X_j à la première direction $\widehat{\alpha}_1$ ainsi que la contribution de X_4, $\widehat{\alpha}_{24}$ à la deuxième direction révélatrice. Les autres estimations $\widehat{\alpha}_{21}$, $\widehat{\alpha}_{22}$, $\widehat{\alpha}_{23}$ des contributions de X_1, X_2, X_3 sont elles plus instables, résultats illustrés par les histogrammes (Figure 1.4 (a) et (b)). Le tableau 1.3 reporte les matrices de corrélations relatives à ces mêmes estimations et révèle de faibles valeurs.

L'exemple met en évidence les caractéristiques de la modélisation PPR; tout d'abord une réduction de la dimension du problème ($p = 4$, $M = 2$), sa flexibilité (prise en compte de liaison non linéaire) et son interprétabilité (favorisée par la représentation graphique des fonctions). La plus grande instabilité du deuxième terme tient à la nature même de l'algorithme PPR, qui utilise les résidus de l'étape 1 pour la détermination des termes suivants. Un biais éventuel de la première direction estimée entraînera une certaine instabilité de la deuxième et ainsi de suite.

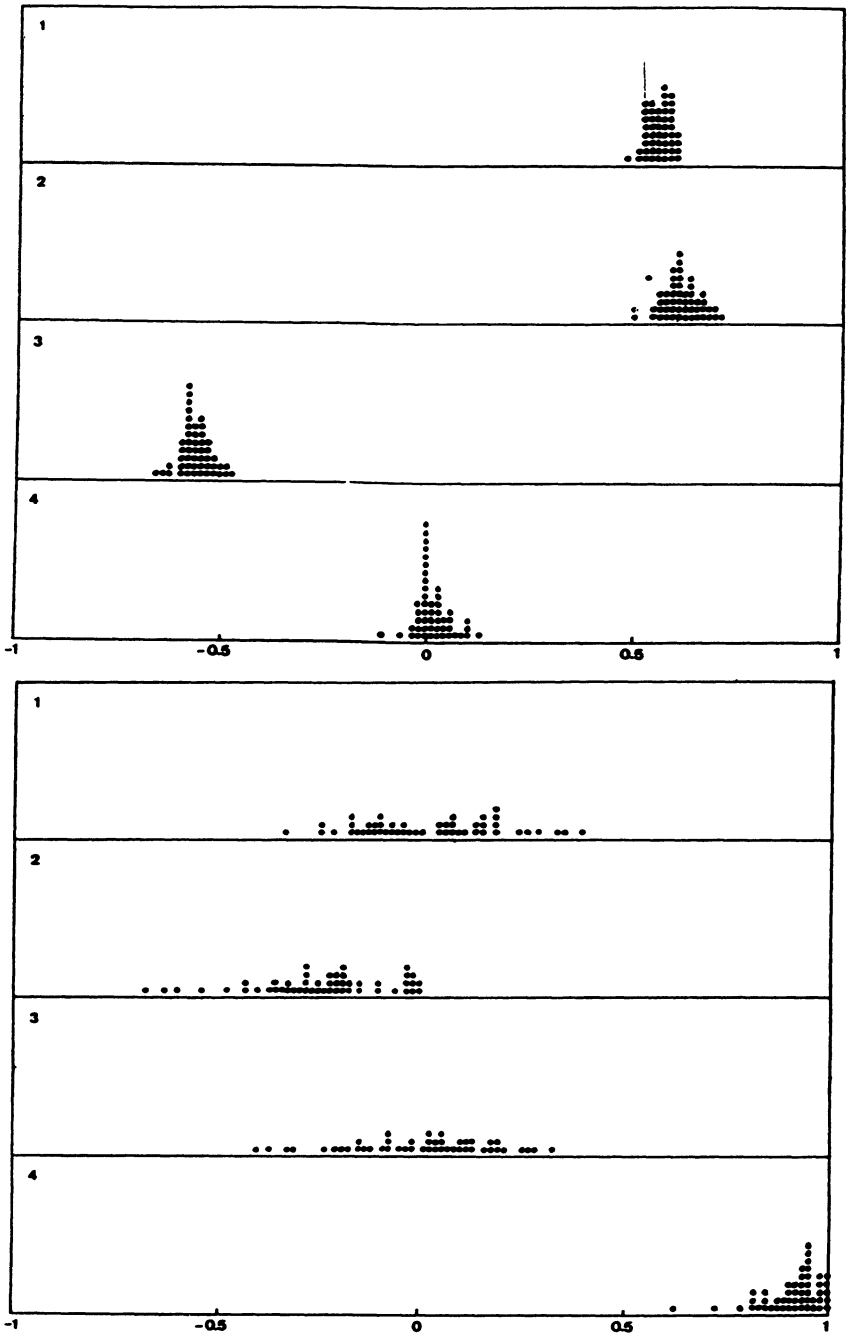
TABLEAU 1.2
Distributions bootstrap des estimateurs des directions $\widehat{\alpha}_1$, $\widehat{\alpha}_2$ (espérance (E^*), écart-type (σ^*))

Paramètres	Estimations	E^*	σ^*
α_{11}	0.550	0.545	0.033
α_{12}	0.586	0.605	0.047
α_{13}	-0.594	-0.573	0.041
α_{14}	-0.010	0.016	0.047
α_{21}	-0.112	0.021	0.179
α_{22}	-0.225	-0.252	0.167
α_{23}	0.200	0.027	0.202
α_{24}	0.947	0.911	0.074

TABLEAU 1.3
Distributions bootstrap des estimateurs des directions $\widehat{\alpha}_1$, $\widehat{\alpha}_2$ (matrices des corrélations)

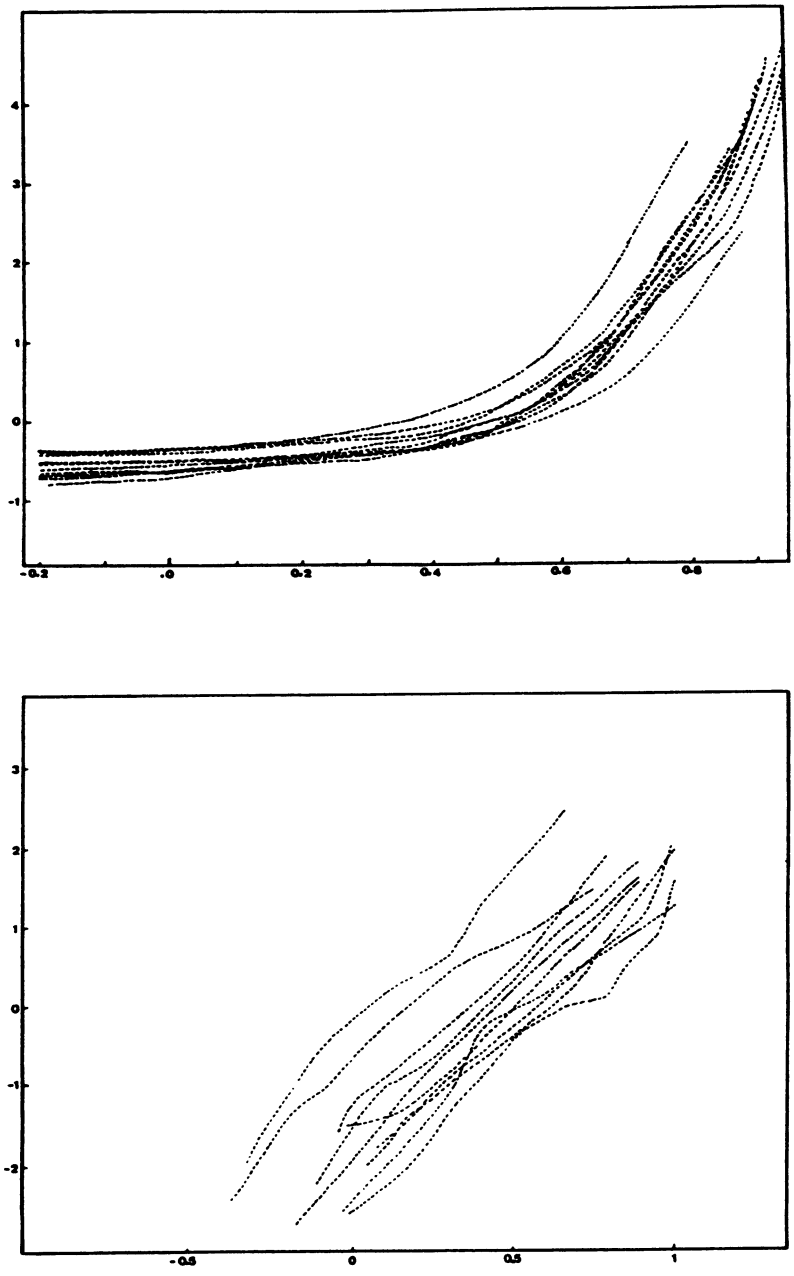
	1	2	3	4
1	1			
2	0.234	1		
3	0.152	0.357	1	
4	0.0003	0.011	0.026	1

	1	2	3	4
1	1			
2	-0.007	1		
3	0.041	0.0	1	
4	0.140	0.122	0.296	1



FIGURES 1.4 (a) et (b)

Histogrammes des distributions bootstrap des estimateurs des deux premières directions.



FIGURES 1.5 (a) et (b)
Représentation graphique des fonctions $\widehat{f}_1(\widehat{\alpha}_1 x)$ *(a) et* $\widehat{f}_2(\alpha_2 x)$ *(b) (10 premiers échantillons bootstrap).*

Les figures 1.5. (a) et (b) représentent les faisceaux de courbes obtenus par bootstrap pour les deux premières fonctions $\widehat{f_1}$ et $\widehat{f_2}$ du développement PPR (nous n'avons représenté que les 10 premières replications bootstrap) et ce conditionnellement à la direction correspondante fixée. Ces fonctions prennent en compte respectivement les parties exponentielle et linéaire du modèle.

Nous avons construit (figure 1.5 (c)) les intervalles de confiance point par point pour la fonction f_1, conditionnellement à la première direction estimée $Z = 0.5(X_1 + X_2 - X_3)$ ainsi que le nuage des observations centrées réduites.

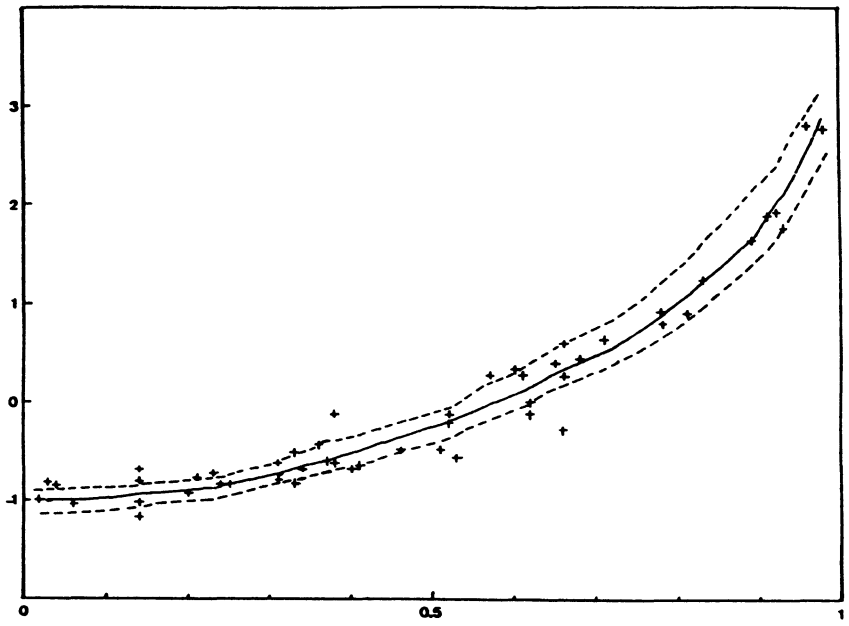


FIGURE 1.5 (c)
Bandes de confiance pour la fonction f_1.

6.2 Simulation 2

Le second exemple consiste en la simulation de $N = 30$ échantillons de taille $n = 50$ observations, générées suivant le modèle :

$$Y = (X_1 + X_2 - X_3) + (X_4 - X_5)^2 + \varepsilon$$

avec $X_1, \ldots, X_5$ v.a. de loi Uniforme $U[0, 1]$ générées indépendamment et $\varepsilon \sim N(0; 0.01)$. Le nombre de replications bootstrap est ici encore de $B = 50$.

La sélection de modèle est ici résumée au tableau 2.1. et conduit au modèle à $M = 3$ termes. Il y a donc désaccord entre modèle identifié et fonction de régression proposée. La raison essentielle est l'existence pour 5 à 6 échantillons de minimas locaux qui piègent l'algorithme dans le processus d'estimation des paramètres. La structure de la fonction de régression n'est pas trouvée et l'introduction de nouveaux

termes ($M = 3$) améliore l'ajustement. Le mauvais comportement de tels échantillons suffit à déplacer le minimum (qui sur les autres échantillons est $M = 2$) sur le modèle à $M = 3$ termes.

Les commentaires du tableau 2.1. sont ici identiques à ceux de l'exemple 1. Les estimations correspondantes aux deux premières directions sont reportées au tableau 2.2. L'importance des variables est ici $1.0(X_1)$, $0.97(X_2)$, $0.89(X_3)$, $0.75(X_4)$ et $0.74(X_5)$.

L'étude des distributions bootstrap ($B = 50$) de ces estimateurs est résumée sur le tableau 2.2 où sont reportées espérance et écart-type correspondants aux coefficients des deux premières directions $\widehat{\alpha}_1, \widehat{\alpha}_2$: on note une relative stabilité des coefficients de la première direction, le deuxième terme étant lui beaucoup plus instable, phénomème également mis en évidence par les histogrammes 2.3.

Les figures 2.4 (a) et (b) représentent les fonctions $\widehat{f}_1$ et $\widehat{f}_2$, d'aspect linéaire et quadratique respectivement. Nous n'avons pas représenté la fonction $\widehat{f}_3$ ni pris en compte la direction révélatrice correspondante, très instables, non interprétables en raison de l'absence de structure définie pour un troisième terme dans l'approximation.

Nous avons, sur cet exemple, étudié la distribution bootstrap du biais de l'erreur de prédiction $\widehat{R}$ sur l'un des échantillons analysés. Les valeurs $R^{*1}, \ldots, R^{*B}$ constituent un échantillon aléatoire, extrait d'une population ayant les caractéristiques suivantes, espérance $E^*(R^*) = \widehat{\omega}_{\text{Boot}} = \widehat{\omega}_\infty$ et variance σ^2. L'histogramme obtenu pour $B = 50$ valeurs (figure 2.5) révèle une distribution voisine d'une normale. Il est possible de donner une mesure de la précision de l'approximation Monte-Carlo $\widehat{r}_B$ de $\widehat{r}_{\text{Boot}}$ pour B échantillons bootstrap, en utilisant l'approximation normale. On déduit alors :

$$|\widehat{\omega}_B - \widehat{\omega}_\infty| \leqslant \frac{2\sigma}{B^{1/2}} \text{ avec probabilité voisine de 1.}$$

Si on estime σ^2 par $\widehat{\sigma}_B^2 = (1/B-1)\sum_{b=1}^{B}(R^{*b} - \widehat{\omega}_B)^2$ on obient pour $B = 50$:

$$|\widehat{\omega}_B - \widehat{\omega}_B| \leqslant 0.01.$$

TABLEAU 2.1

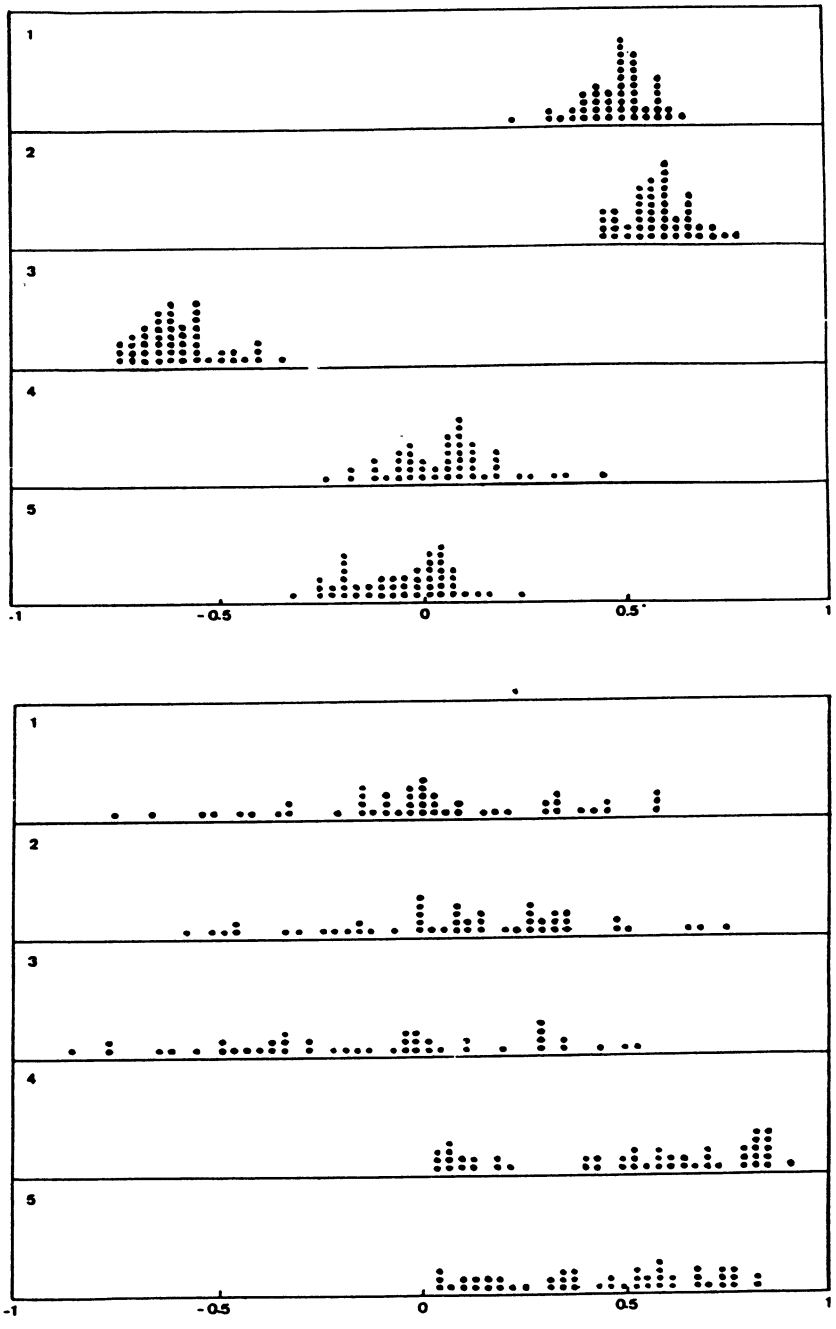
Critères $\widehat{r}_{AE}$, $\widehat{r}_{CV}$, $\widehat{r}_{B}$ et estimation $\widehat{r}$ de l'erreur de prédiction, espérance (E), écart-type (σ).

Critères		M=1	M=2	M=3	M=4
$\hat{r}_{AE}$	(E)	0.0301	0.0090	0.0025	0.0018
	(σ)	0.0068	0.0080	0.0048	0.0038
$\hat{r}_{CV}$	(E)	0.0530	0.0343	0.0226	0.0237
	(σ)	0.0116	0.0091	0.0094	0.0081
$\hat{r}_{B}$	(E)	0.0581	0.0306	0.0212	0.0217
	(σ)	0.0146	0.0099	0.0080	0.0077
$\hat{r}$	(E)	0.0481	0.0400	0.0199	0.0200
	(σ)	0.0152	0.0190	0.0216	0.0208

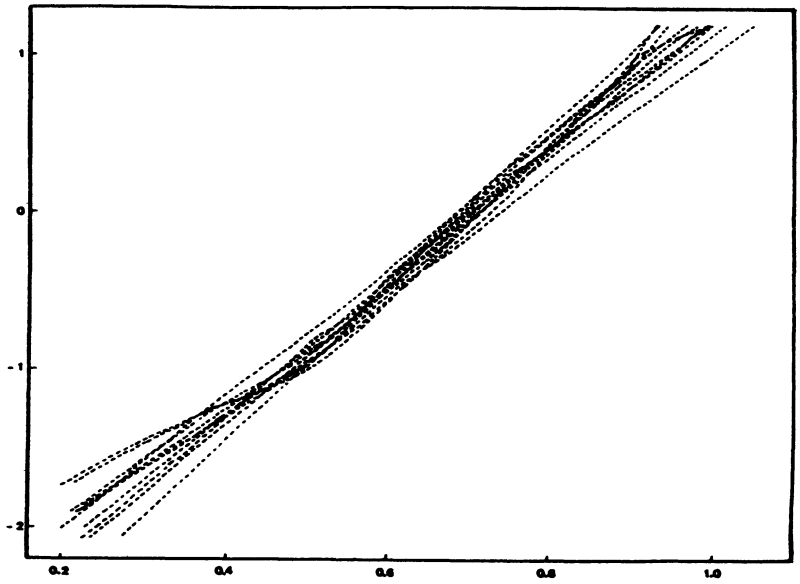
TABLEAU 2.2

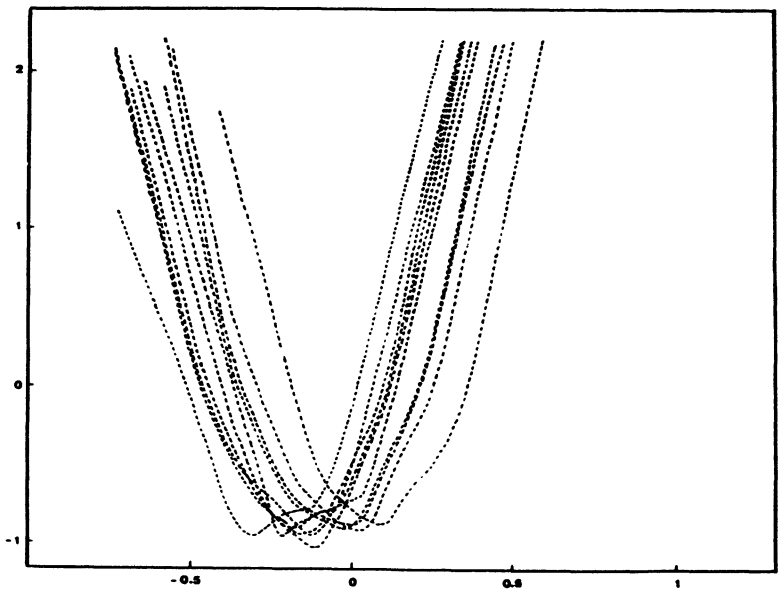
Distribution bootstrap des estimateurs des directions $\widehat{\alpha}_1$, $\widehat{\alpha}_2$.

Paramètres	Estimations	E^*	σ^*
α_{11}	0.575	0.487	0.084
α_{12}	0.581	0.581	0.081
α_{13}	-0.574	-0.597	0.095
α_{14}	-0.028	0.046	0.138
α_{15}	0.040	0.075	0.125
α_{21}	-0.01	0.014	0.331
α_{22}	-0.01	0.085	0.318
α_{23}	-0.003	-0.036	0.442
α_{24}	0.C8	0.501	0.291
α_{25}	-0.73	0.443	0.249



FIGURES 2.3 (a) et (b)
Histogrammes des distributions bootstrap des estimateurs des deux premières directions.





FIGURES 2.4 (a) et (b)
Représentation graphique des fonctions $\widehat{f}_1(\widehat{\alpha}_1 x)$ *(a) et* $\widehat{f}_2(\widehat{\alpha}_2 x)$ *(b) (10 premiers échantillons bootstrap).*

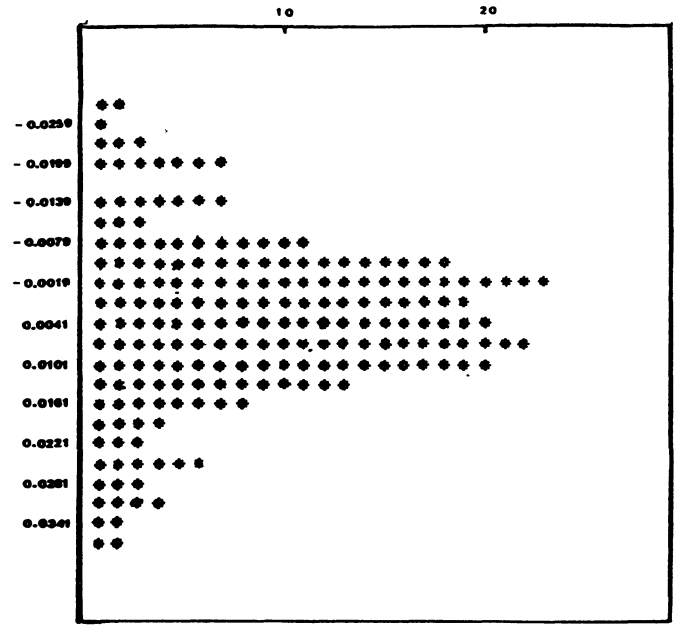


FIGURE 2.5
Critère bootstrap, distribution Monte-Carlo de l'erreur $R(F, F)$ *avec* $B = 50$ *échantillons.*

7. Discussion

Les simulations réalisées ont permis de préciser les performances de la régression par directions révélatrices et les problèmes qu'elle soulève. La flexibilité, l'interprétabilité de ces modèles ont été mises en évidence. Il n'est pas inutile à ce propos de rappeler que les méthodes PPR s'insèrent dans un contexte d'analyse exploratoire donc dans des situations ou l'on ne dispose pas d'information *a priori* sur la fonction de régression (ce qui exclut un modèle paramétrique) et où la dimension du problème est élevée ($p \geqslant 10$) (ce qui exclut une extension des techniques classiques de régression non paramétrique, qui ne fonctionnent correctement que si $p \geqslant 3$). Il n'y a donc pas dans ce type de situations d'alternatives (ou de concurrents) directs.

– Un des premiers problèmes soulevés est celui au choix d'un modèle, dans la hiérarchie de modèles de complexité croissante proposée. Les simulations ont fait ressortir trois types de difficultés : tout d'abord la très grande sensibilité de l'algorithme d'estimation aux minimums locaux, dans la phase de recherche des directions. Une deuxième difficulté est le risque de surajustement de la procédure «supersmoother». Enfin il faut mentionner les temps de calcul très importants mis en jeu pour une telle sélection par rééchantillonnage.

– Un second problème est celui de l'inférence, difficile à développer dans le contexte des modèles PPR complets sinon par rééchantillonnage. La difficulté majeure est constituée, dans ce domaine, par le biais possible dans la recherche des directions.

Une solution a été proposée par Hall (1989), qui consiste en une recherche s'articulant sur deux étapes. La première étape est une estimation sous lissée permettant de déterminer la bonne orientation pour la direction cherchée. Le deuxième niveau consiste à reconstruire l'estimation de la fonction de régression avec un degré de lissage correct. Une telle procédure devrait permettre d'améliorer la convergence de l'estimation des directions et ne pas affecter l'estimation des fonctions. Toutefois le lissage de la première étape rend le problème des minimas locaux plus aigü.

Différents auteurs ont également proposé certaines solutions originales au problème de l'inférence pour un modèle de type PPR. C'est le cas de la méthode baptisée PPR-Adjoint (Duan, 1990), de la méthode SIR (Sliced Inverse Regression) (Ker-Chau-Li, 1991) ou encore de la méthode ADE (Average Derivative Estimation), Härdle (1990).

– Un troisième problème soulevé est celui du type de fonctions de régression pour lesquelles une approximation de type PPR est bien adaptée. L'existence de fonctions simples, nécessitant un développement avec un nombre de termes élevé en est une illustration (Diaconis et Shashahani, 1984). Cet aspect a été étudié par Donoho et Johnstone (1989), qui ont comparé les performances de PPR avec une méthode à noyau multidimensionnelle dans le cas de fonctions de $\mathbb{R}^2$. Dans le cas général il semble que PPR puisse modéliser des fonctions de structure additive mais rencontre des difficultés pour certaines fonctions comportant des interactions. PPR peut prendre en compte certaines interactions : un exemple simple en est la fonction $f(X_1, X_2) = X_1 X_2$ (Friedman, Stuetzle, 1981) qui peut se réécrire sous la forme :

$$f(X_1, X_2) = (1/4)(X_1 + X_2)^2 - (1/4)(X_1 - X_2)^2.$$

Toutefois pour des interactions plus complexes, le développement risque d'être peu interprétable. Dans ce cas, les méthodes dites de «régression par arbres» comme la méthode CART (Breiman, Friedman, Olshen, Stone, 1984) ou plus récemment la méthode MARS («Multivariate Adaptative Regression Splines») (Friedman, 1991) semblent prometteuses et présenter une certaine complémentarité avec une approche additive de type PPR.

En conclusion, la régression par directions révélatrices semble devoir être développée dans une phase exploratoire d'une analyse de régression, pour mettre en évidence d'éventuelles structures sous jacentes. Les modèles PPR peuvent alors servir d'indicateurs, pour effectuer certaines transformations des variables initiales, permettant ensuite l'étude de modèles paramétriques.

Remerciements

Les auteurs remercient les relecteurs pour leurs suggestions, qui ont permis d'améliorer la forme définitive de l'article.

Bibliographie

(1) BREIMAN L., FRIEDMAN J.H. (1985), *Estimating optimal transformation for multiple regression and correlation*, J. of the Amer. Stat. Assoc., 80, p. 580-619.

(2) BREIMAN L., FRIEDMAN J.H., OLSHEN R., STONE C.J. (1984), *Classification and regression trees*, Wadworth, Belmont.

(3) BUJA A., HASTIE T., TIBSHIRANI R. (1989), *Linear smoothers and additive models*, Annals of Statistics, 17, p. 453-555.

(4) BUNKE O., DROGE B. (1984a), *Estimators of the mean squared error of prediction in linear regression*, Technometrics,26, p. 145-155.

(5) BUNKE O., DROGE B. (1984b), *Bootstrap and cross-validation estimates of the prediction error for linear regression models*, Annals of Statistics, 12, p. 1400-1424.

(6) COLLOMB G. (1981), *Estimation non paramétrique de la regression : revue bibliographique*, Int. Stat. Rev.,49, p. 75-93.

(7) DAUDIN J.J., LEGER L. (1989), *Erreur de prédiction et sélection de modèles en Projection Pursuit Regression*, Communications ISI Paris.

(8) DIACONIS P., SHASHAHANI M. (1984), *On non linear functions of linear combinations*, SIAM, J. Sci. Statist. Comput., 5, p. 175-191.

(9) DONOHO D.L., JONHSTONE I.M. (1989), Projection based approximations and a duality with kernel metods, Annals of Statistics, 17, p. 58-106.

(10) DROGE B. (1987), *A note on estimating the M.S.E.P. in non linear regression*, Statistics, 18, p. 499-520.

(11) DUAN N. (1990), *The adjoint projection pursuit regression*, Annals of Statistics, 85, p 1029-1038.

(12) EFRON B. (1979), *Bootstrap methods : another look at the jackknife*, Annals of Statistics, 7, p. 1-26.

(13) EFRON B., (1983a), *The jackknife, the bootstrap, and other resampling plans*, SIAM, n 38, Philadelphia.

(14) EFRON B. (1983b), *Estimating the error rate of a prediction rule : improvements of cross-validation*, J. of the Amer.Stat. Assoc., 78, p. 316-331.

(15) EFRON B. (1988), *Computer intensive metods in statistical regression*, SIAM Review, 30, p. 421-449.

(16) EFRON B., TIBSHIRANI R. (1986), *Bootstrap metods for standard errors, confidence intervals, and others measures of statistical accuracy*, Stat. Science, 1, p. 54-77.

(17) EUBANK R. L. (1988), *Spline smoothing and non parametric regression*, Marcel Dekker, New-York, 1988.

(18) FREEDMAN D. (1981), *Bootstraping regression models, Annals of Statistics*, 9, p. 1218-1228.

(19) FRIEDMAN J.H. (1984 a), *Smart users'guide*, Tech. Report LCM001 Dept of Statistics, Stanford University.

(20) FRIEDMAN J.H. (1984 b), *Classification and multiple response regression through projection pursuit*, Tech. Report LCM006, Dept of Statistics, Stanford University.

(21) FRIEDMAN J.H. (1991), *Multivariate adaptative regression splines*, Annals of Statistics, p. 1-115.

(22) FRIEDMAN J.H., GROSSE E., STUETZLE W. (1983), *Multidimensional splines approximation*, SIAM, J. of Sci. Stat. Comp., p. 291-301.

(23) FRIEDMAN J.H., SILVERMAN B.W. (1989), *Flexible parsimonious smoothing and additive modelling*, Technometrics, 31, p. 3-29.

(24) FRIEDMAN J.H., STUETZLE W. (1981), *Projection pursuit regression J. Amer. Stat. Soc.*, 76, p. 817-823.

(25) FRIEDMAN J.H., STUETZLE W., SCHROEDER A. (1984), *Projection pursuit density estimation*, J. Amer. Stat. Soc., 76, p. 599-608.

(26) FRIEDMAN J.H., TUCKEY J.W. (1974), *A projection pursuit algorithm for exploratory data analysis*, IEEE Trans. Comput. 23, p. 881-889.

(27) HALL P. (1989), *On projection pursuit regression.* Annals of Statistics 17, p. 537-588.

(28) HARDLE W. (1990), *Applied non parametric regression*, Cambridge Univ. Press.

(29) HASTIE T., TIBSHIRANI R. (1986), *Generalized additive models*, Stat. Science, 1 , p. 297-318.

(30) HASTIE T., TIBSHIRANI R. (1987), *Generalized additive models : some applications*. J. of the Amer. Stat. Assoc., 82 , p. 371-386.

(31) HASTIE T., TIBSHIRANI R. (1990), *Generalized additive models*, Chapman-Hall.

(32) HOCKING R. (1976), *The analysis and selection of variables in linear regression*, Biometrics, 32, p. 1-49.

(33) HUBER P.J. (1985), *Projection pursuit*, Annals of Statistics, 13 p. 435-525.

(34) JOHANSEN S., JOHNSTONE I.M. (1990), *Hotteling's theorem on the volume of tubes : some illustrations in simultaneous inference and data analysis*, Annals of Statistics, 18, p. 652-684.

(35) KER CHAU LI (1991), *Sliced inverse regression for dimension reduction*, Annals of Statistics, 86, p. 316-342.

(36) LEGER L. (1991), *Régression par la méthode des directions révélatrices : présentation générale et sélection de modèles*, Thèse de Doctorat, INA-PG,

(37) LINHART H., ZUCCHINI O.(1985), *Model selection*, wiley, New-York.

(38) ROSENBROCK H.H. (1960), *An automatic method for finding the greatest or least value of a function*, Comp. J. , 3, p. 175-184.

(39) SILVERMAN B.W. (1985), *Some aspects of the spline smoothing approach to non parametric regression curve fitting*, J. of the Roy. Stat. Soc., Ser B, 47, p. 1-52.

(40) STONE C.J. (1985), *Additive regression and other non parametric model*, Annals of Statistics, 13, p. 689-705.

(41) STONE M. (1974), *Cross validatory choice and assesment of statistical prediction*, J. of the Roy. Stat. Soc., Ser B, p. 111-147.

(42) TOMASSONE R., DAUDIN J.J., DANZART M., MASSON J.P. (1988), *Discrimination et classement*, Masson, Paris.