Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Acad. Sci. Paris, Ser. I 341 (2005) 365–368
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistique

Sélection automatique du paramètre de lissage pour l'estimation non paramétrique de la régression pour des données fonctionnelles

Mustapha Rachdi [a], Philippe Vieu [b]

[a] *Université Pierre Mendès France, UFR SHS, BP. 47, 38040 Grenoble cedex 09, France*
[b] *Université Paul Sabatier, LSP UMR CNRS 5583, 118, route de Narbonne, 31062 Toulouse cedex, France*

Reçu le 7 octobre 2004 ; accepté après révision le 20 juin 2005

Présenté par Paul Deheuvels

Résumé

Dans cette Note, nous étudions l'estimation de la régression quand le régresseur est de type fonctionnel. L'estimateur de la régression pour ce type de données a été récemment introduit. Il dépend d'un paramètre de lissage qui contrôle sa vitesse de convergence, et le but de notre travail est de construire un critère de choix automatique de ce paramètre. Le critère est formulé sous la forme d'une validation croisée fonctionnelle. Sous certaines hypothèses sur l'opérateur de regression (inconnu), nous montrons que cette procédure est optimale. En plus, nous établissons l'équivalence asymptotique entre plusieurs mesures de risque pour l'estimation non paramétrique de l'opérateur de régression. ***Pour citer cet article : M. Rachdi, P. Vieu, C. R. Acad. Sci. Paris, Ser. I 341 (2005).***

Abstract

Automatic smoothing parameter selection for the nonparametric regression estimation of functional data. We study regression estimation when the explanatory variable is functional. Nonparametric estimates of the regression operator have been recently introduced. They depend on a smoothing factor which controls its behaviour, and the aim of our Note is to construct some data-driven criterion for choosing this smoothing parameter. The criterion can be formulated in terms of a functional version of cross-validation ideas. Under mild assumptions on the unknown regression operator, it is seen that this rule is asymptotically optimal. As by-products of this result, we state asymptotic equivalences for several measures of accuracy for nonparametric estimate of the regression operator. ***To cite this article: M. Rachdi, P. Vieu, C. R. Acad. Sci. Paris, Ser. I 341 (2005).***

Adresses e-mail : mrachdi@upmf-grenoble.fr (M. Rachdi), vieu@cict.fr (P. Vieu).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2005.06.027

1. Introduction

Les données fonctionnelles ont fait leur apparition dans plusieurs domaines de la statistique appliquée (médecine, environnement, ...). Il est donc de plus en plus fréquent de travailler avec ce type de données. D'un point de vue technique, un échantillon de données fonctionnelles peut être rencontré dans beaucoup de problèmes statistiques (classification, discrimination, études longitudinales, prévision, ...). Ce champ de la statistique moderne est devenu populaire grâce au livre de Ramsay et Silverman [5] et est généralement connu sous le nom d'*Analyse Statistique des Données Fonctionnelles*. Le lecteur peut trouver dans Ferraty et Vieu [2] une vue d'ensemble sur des problématiques et des avancées récentes liées à ce domaine important de la statistique moderne.

Dans cette Note, nous nous intéressons à la prévision d'une variable réponse scalaire Y étant donnée une certaine variable fonctionnelle X. En d'autres termes, la question est d'estimer $r(\cdot) = \mathbb{E}(Y|X = \cdot)$ quand X est une variable aléatoire à valeurs dans un espace de dimension éventuellement finie. L'estimation de l'opérateur r a été traitée par plusieurs auteurs durant la dernière décennie, et la plupart de ces travaux concernaient les modèles linéaires pour l'opérateur r (voir Cardot et al. [1]). En suivant les travaux de Ferraty et Vieu [2], une approche non paramétrique du problème est envisageable, mais comme c'est le cas avec plusieurs estimateurs non paramétriques, aussi bien dans le cadre non fonctionnel que dans le cas de dimension infinie étudié ici, il y a un paramètre de lissage qui intervient dans la construction des estimateurs et qui doit être sélectionné convenablement afin d'assurer de bonnes performances sur le plan pratique. Le but principal de cette note est donc de proposer une procédure automatique pour choisir ce paramètre, et d'établir son optimalité asymptotique dans un sens quadratique.

Cette Note est organisée comme suit. La Section 2 est consacré à la présentation du modèle, à l'estimation des opérateurs et à la description de la procédure de choix du paramètre de lissage. La Section 3 contient le résultat principal de cette Note, à savoir le Théorème 3.1 qui est une extension au cas de régresseurs de dimension infinie des résultats de Härdle et Marron [3] et Marron et Härdle [4], et contient également des résultats d'équivalences asymptotiques entre différentes mesures de risque quadratique pour les variables fonctionnelles.

2. Critère de choix de la largeur de fenêtre

Soit (E, D) un espace semi-métrique (dimension infinie). Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ un échantillon de réalisations de la variable aléatoire (X, Y) à valeurs dans $E \times \mathbb{R}$. Suivant Ferraty et Vieu [2], le problème d'estimation de l'opérateur fonctionnel r peut être traité par les techniques d'estimation à noyau. Précisément, ces auteurs ont introduit l'estimateur opératoriel suivant :

$$\hat{r}_h(x) = \frac{\sum_{i=1}^n Y_i K(h^{-1} D(x, X_i))}{\sum_{i=1}^n K(h^{-1} D(x, X_i))}$$

où K est une fonction réelle définie sur $\mathbb{R}^+$ et $h = h(n) \in \mathbb{R}^+$ est le paramètre de lissage (largeur de fenêtre). Pour la suite, nous rappelons l'erreur quadratique moyenne et l'erreur quadratique moyenne integrée : $ASE(h) = n^{-1} \sum_{j=1}^n (\hat{r}_h(X_j) - r(X_j))^2 W(X_j)$ et $MISE(h) = \int \mathbb{E}(\hat{r}_h(x) - r(x))^2 W(x)\, \mathrm{d}F(x)$, où W est une fonction de poids positive (connue) et $\mathrm{d}F$ désigne la mesure de répartition.

En s'inspirant de Härdle et Marron [3], nous introduisons le critère :

$$CV(h) = n^{-1} \sum_{j=1}^n \left(Y_j - \hat{r}_h^j(X_j)\right)^2 W(X_j)$$

où $\hat{r}_h^i(x)$ est l'estimateur *validé croisé* de $r(x)$ donné par : $\hat{r}_h^j(x) = \dfrac{\sum_{\substack{i=1 \\ i \neq j}}^n Y_i K(h^{-1} D(x, X_i))}{\sum_{\substack{i=1 \\ i \neq j}}^n K(h^{-1} D(x, X_i))}$.

Le critère CV est connu sous le nom de *critère de validation croisée* et son minimiseur $h_0 = \arg\min_{h \in H_n} CV(h)$ est appelé *la largeur de fenêtre validée croisée*.

3. Optimalité asymptotique du paramètre de lissage sélectionné

3.1. Hypothèses

Fonction de poids. Nous supposons que :

$$W \text{ est une fonction positive bornée et à support compact d'intérieur non vide.} \tag{1}$$

Largeur de fenêtre. Nous supposons que :

$$H_n \text{ est un ensemble fini de paramètres } h, \text{ tels que } h \to 0 \text{ et satisfaisant la condition (5).} \tag{2}$$

Le noyau. Nous supposons que K est strictement décroissant sur $[0,1]$ et $\exists a, b > 0$ tels que :

$$a\,\mathbb{1}_{[0,1]}(x) \leqslant K(x) \leqslant b\,\mathbb{1}_{[0,1]}(x), \quad \text{pour tout } x \in \mathbb{R}. \tag{3}$$

Concentration de X. Nous supposons que la loi de probabilité de la variable fonctionnelle X peut être écrite comme suit

$$\mathbb{P}\big(D(x, X_i) \leqslant h\big) = C_x \varphi(h) + o\big(\varphi(h)\big), \quad \text{pour tout } x \in E,\ i \in \{1, \ldots, n\}, \tag{4}$$

et satisfaisant, pour tout compact $\mathcal{S}$, les conditions : $\sup_{x\in\mathcal{S}} C_x < \infty$, $\varphi(s) > 0$, $\forall s$, $\lim_{s\to 0} \varphi(s) = 0$, et

$$\exists \tau > 0 \text{ tel que } \sum_{n=1}^{+\infty} \varphi(h)^{\tau} < \infty. \tag{5}$$

Opérateur de régression. Le modèle statistique pour l'opérateur fonctionnel r est non paramétrique en ce sens que nous supposons l'hypothèse de régularité suivante réalisée :

$$\exists C < \infty,\ \exists \beta > 0,\ \text{tel que } \forall x, y \in E: \quad \big|r(x) - r(y)\big| \leqslant C D(x, y)^{\beta}. \tag{6}$$

Moments conditionnels. Nous supposons l'hypothèse de bornage habituelle suivante :

$$\forall k \in \mathbb{N}^*,\ \exists C_k > 0 \text{ tel que } \mathbb{E}\big(|Y|^k | X = x\big) \leqslant C_k \text{ pour tout } x \in E. \tag{7}$$

$$\exists \sigma > 0 \text{ tel que } \mathbb{E}\big(Y^2 | X = x\big) = \sigma(x) \geqslant \sigma > 0 \text{ pour tout } x \in E. \tag{8}$$

Notre jeu d'hypothèses est relativement peu restrictif, puisque les conditions sur Y, W K et H_n sont les mêmes qu'en dimension finie (voir Härdle et Marron [3]) tandis que la condition (4) sur X est la même que celle nécessaire pour obtenir la convergence des estimateurs (voir Ferraty et Vieu [2]).

3.2. Résultat principal

Nous sommes maintenant en mesure d'énoncer le résultat principal de cette note.

Théorème 3.1. *Sous les hypothèses* (1)–(8)*, le critère de sélection du paramètre de lissage qui consiste à choisir $h_0 \in H_n$ minimisant $CV(h)$ est asymptotiquement optimal relativement aux distances $d =$ ASE ou MISE, au sens suivant* :

$$\frac{d(\hat{r}_{h_0}, r)}{\inf_{h\in H_n} d(\hat{r}_h, r)} \to 1, \quad \textit{p.s. quand } n \to +\infty.$$

Idées de la démonstration. Ce résultat se démontre en suivant le même cheminement qu'en dimension finie (voir Härdle et Marron [3]), mais en tenant compte de l'aspect fonctionnel des variables grâce à des techniques similaires à celles employées par Ferraty et Vieu [2]. Les deux outils clés employés dans cette preuve sont des équivalences asymptotiques entre divers critères d'erreurs quadratiques (que nous établirons et que nous rappelons au paragraphe suivant), ainsi que des inégalités de moments. L'intégralité de nos preuves peut être obtenue sur demande. □

3.3. Équivalence entre les erreurs quadratiques

Nous établissons trois lemmes concernant l'équivalence asymptotique entre plusieurs mesures de risque. Ces résultats jouent un rôle important dans la preuve de notre principal résultat.

Lemme 3.2. *Sous les hypothèses* (2), (3), (5)–(7) *et* (8), *nous avons*

$$\sup_{h\in H_n}\left|\frac{MISE(h)-ISE(h)}{MISE(h)}\right|\to 0,\quad \textit{p.s. quand } n\to\infty.$$

Lemme 3.3. *Sous les hypothèses du Lemme* 3.2, *et si en plus une des trois conditions suivantes est satisfaite* : (i) nh^k *est bornée, pour* $k\geqslant 2$, *ou* (ii) $\varphi(h)n^{1/2}h^{2\beta}<1$, *ou* (iii) $h\leqslant Cn^{-\kappa}$ *pour* $\kappa\geqslant 1/4\beta$, *alors, on a*

$$\sup_{h\in H_n}\left|\frac{MISE(h)-ASE(h)}{MISE(h)}\right|\to 0,\quad \textit{p.s. quand } n\to\infty.$$

Lemme 3.4. *Sous les hypothèses* (2), (3), (5)–(7) *et* (8), *nous avons*

$$\sup_{h\in H_n}\left|\frac{ASE(h)-ISE(h)}{ISE(h)}\right|\to 0,\quad \textit{p.s. quand } n\to\infty.$$

Remarque 1. Des Lemmes 3.2, 3.3 et 3.4 on peut établir des résultats généraux semblables à ceux démontrés par Vieu [6] dans le cas réel :

$$\sup_{h\in H_n}\left|\frac{d(\hat r_h(x),r(x))-d'(\hat r_h(x),r(x))}{d(\hat r_h(x),r(x))}\right|\to 0,\quad \text{p.s. quand } n\to\infty$$

où $d, d'\in\{ASE, ISE, MISE\}$.

Remerciements

Nous tenons à remercier F. Ferraty. Ses connaissances en Statistique non paramétrique pour Données Fonctionnelles et ses remarques pertinentes, sur le plan théorique et pratique, ont été particulièrement bénéfiques. Nous adressons également nos remerciements aux participants du groupe STAPH (Statistique Fonctionnelle et Operatoriel de Toulouse), pour leurs remarques et commentaires qui furent fructueux.

Références

[1] H. Cardot, F. Ferraty, P. Sarda, Spline estimators for the functional linear model, Statist. Sinica 13 (3) (2003) 571–591.
[2] F. Ferraty, P. Vieu, Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination, J. Nonparametr. Statist. 16 (1–2) (2004) 111–125.
[3] W. Härdle, J.S. Marron, Optimal bandwidth selection in nonparametric regression function estimation, Ann. Statist. 13 (4) (1985) 1465–1481.
[4] J.S. Marron, W. Härdle, Random approximations to some measures of accuracy in nonparametric curve estimation, J. Multivariate Anal. 20 (1986) 91–113.
[5] J. Ramsay, B. Silverman, Functional Data Analysis, Springer-Verlag, 1997.
[6] P. Vieu, Quadratic errors for nonparametric estimates under dependence, J. Multivariate Anal. 39 (2) (1991) 324–347.