



ELSEVIER

Contents lists available at ScienceDirect

C. R. Acad. Sci. Paris, Ser. I

www.sciencedirect.com



Statistique

Estimation locale linéaire de la fonction de régression pour des variables hilbertiennes



Local linear estimation of the regression function with Hilbertian variables

Jacques Demongeot^a, Ali Laksaci^b, Amina Naceri^b, Mustapha Rachdi^c

^a Université Grenoble Alpes, laboratoire AGEIS EA 7407, faculté de médecine de Grenoble, 38700 La Tronche, France

^b Laboratoire de statistique et processus stochastiques, université Djillali-Liabès, Sidi Bel-Abbès, BP 89, Sidi Bel-Abbès, 22000, Algérie

^c Université Grenoble Alpes, laboratoire AGEIS EA 7407, UFR SHS, BP 47, 38040 Grenoble cedex 09, France

IN F O A R T I C L E

Historique de l'article :

Reçu le 13 décembre 2015

Accepté après révision le 7 juin 2016

Disponible sur Internet le 20 juillet 2016

Présenté par Paul Deheuvels

R É S U M É

Dans cette Note, nous étudions l'estimation non paramétrique de la fonction de régression, lorsque la variable réponse et la covariable sont fonctionnelles. Nous construisons un estimateur local linéaire de l'opérateur de régression, et nous évaluons son erreur d'estimation. Ensuite, nous démontrons sa convergence presque complète et uniforme. La vitesse de convergence obtenue est exprimée en fonction de la probabilité des petites boules de la covariable et en fonction de la fonction d'entropie de l'ensemble sur lequel la convergence uniforme est obtenue.

© 2016 Académie des sciences. Publié par Elsevier Masson SAS. Cet article est publié en Open Access sous licence CC BY-NC-ND (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

A B S T R A C T

In this paper, we introduce a new nonparametric estimation of the regression function when both the response and the explanatory variables are of the functional kind. First, we construct a local linear estimator of this regression operator, then we state its rate for the uniform almost complete convergence. This latter is expressed as a function of the small ball probability of the predictor and as a function of the entropy of the set on which the uniformity is obtained.

© 2016 Académie des sciences. Publié par Elsevier Masson SAS. Cet article est publié en Open Access sous licence CC BY-NC-ND (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Un problème récurrent en statistiques est celui où l'on cherche à expliquer comment se comporte une variable d'intérêt Y en fonction d'une variable explicative X . Dans cette note, nous nous proposons d'étudier ce lien lorsque les deux variables sont hilbertiennes (à valeurs dans des espaces de Hilbert, notés respectivement \mathcal{F} et \mathcal{H}) :

Adresses e-mail : jacques.demongeot@agim.eu (J. Demongeot), alilak@yahoo.fr (A. Laksaci, A. Naceri), mustapha.rachdi@univ-grenoble-alpes.fr, mustapha.rachdi@agim.eu (M. Rachdi).

<http://dx.doi.org/10.1016/j.crma.2016.05.017>

1631-073X/© 2016 Académie des sciences. Publié par Elsevier Masson SAS. Cet article est publié en Open Access sous licence CC BY-NC-ND (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

$$Y = r(X) + \varepsilon, \quad (1)$$

où r est un opérateur de \mathcal{F} dans \mathcal{H} et ε est une variable aléatoire d'erreur.

La modélisation statistique des données fonctionnelles a connu un grand essor. Ce dernier est dû à la diversité des champs d'application dans lesquels les observations se présentent sous forme de courbes, de surfaces ou d'images (cf. [16] pour une référence de base et [6,14,15,17] pour un accès aux références les plus récentes).

Rappelons que le modèle de régression, qu'il soit paramétrique ou non paramétrique, reste l'outil le plus privilégié pour analyser la co-variabilité entre X et Y (quand seule la covariable est fonctionnelle, cf. [5] et [16] pour le cadre paramétrique et [13] pour le cadre non paramétrique). Notons que la littérature, sur le modèle non paramétrique de régression lorsque les deux variables sont fonctionnelles, reste très restreinte. En effet, on peut citer [7], qui traite de la convergence en norme L_p de l'estimateur à noyau de la régression non paramétrique, ainsi que [10], qui traite de la convergence presque complète et uniforme de l'estimateur à noyau de la fonction de régression d'une variable banachique sachant une variable explicative à valeurs dans un espace semi-métrique. En outre, la généralisation de ce résultat au cas de données dépendantes a été obtenue dans [11] (cf. aussi [12] pour la normalité asymptotique).

Dans la plupart des travaux sus-cités, les auteurs ont adopté la méthode à noyau pour l'estimation de l'opérateur de régression. Cependant, il est bien connu que cette méthode produit un biais plus élevé comparée à la méthode locale linéaire (cf. [9] pour le cas d'une covariable non fonctionnelle). Rappelons que cette approche a été introduite en analyse statistique des données fonctionnelles dans [1–4,8].

Notre objectif est donc de généraliser [1] au cas où la variable réponse est également de type fonctionnel. Dans cette direction, nous nous attelons à établir la convergence presque complète et uniforme de l'estimateur local linéaire construit. La vitesse de convergence, que nous avons obtenue, confirme la supériorité de l'estimateur local linéaire sur l'estimateur à noyau, car le premier présente un biais moins important.

Nous présentons notre modèle et son estimateur dans le paragraphe suivant. Dans le troisième paragraphe, nous donnons une évaluation de l'erreur d'estimation. Notre résultat principal, ainsi qu'une esquisse de sa preuve, sont donnés dans le paragraphe 4. Le dernier paragraphe est consacré à une discussion conclusive.

2. Le modèle et son estimateur

Soit (X, Y) un couple de variables aléatoires à valeurs dans $\mathcal{F} \times \mathcal{H}$, où $(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mathcal{F}})$ et $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ sont deux espaces de Hilbert. Nous notons par $\|\cdot\|_{\mathcal{F}}$ (resp. $\|\cdot\|_{\mathcal{H}}$) la norme induite du produit scalaire $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ (resp. $\langle \cdot, \cdot \rangle_{\mathcal{H}}$). Nous considérons un échantillon constitué de n paires indépendantes (X_i, Y_i) qui suivent la même loi que (X, Y) .

Généralement, l'estimation de r par polynômes locaux est fondée sur une condition de régularité permettant d'approximer localement l'opérateur de régression par un polynôme. Afin de construire l'estimateur local linéaire de l'opérateur r , nous procédons par l'approximation de celui-ci :

$$r(x_0) = a(x) + b_x(x_0 - x) + \rho_x(x_0 - x, x_0 - x) + o(\|x_0 - x\|_{\mathcal{F}}^2), \quad (2)$$

pour tout x_0 dans un voisinage de x de \mathcal{F} où b_x est un opérateur linéaire (resp. ρ_x est un opérateur bilinéaire) de \mathcal{F} vers \mathcal{H} (resp. de $\mathcal{F} \times \mathcal{F}$ vers \mathcal{H}). Les deux opérateurs a et b_x sont à estimer en cherchant la solution du problème de minimisation :

$$\min_{a,b} \sum_{i=1}^n \|Y_i - a - b(X_i - x)\|_{\mathcal{H}}^2 K(h^{-1}\|x - X_i\|_{\mathcal{F}}) \quad (3)$$

où K est un noyau et $h = h_{K,n}$ est une suite de nombres réels positifs. En utilisant la linéarité de b_x et la décomposition de $(X_i - x)$ sur une base orthonormée $(v_j)_{j \geq 1}$ de \mathcal{F} , nous remplaçons le critère (3) par :

$$\min_{a,b} \sum_{i=1}^n \left\| Y_i - a - \sum_{j \geq 1} c_{ij} b_x(v_j) \right\|_{\mathcal{H}}^2 K(h^{-1}\|x - X_i\|_{\mathcal{F}})$$

où c_{ij} sont les coefficients de $(X_i - x)$ dans la base $(v_j)_{j \geq 1}$. Comme ce critère n'est pas exploitable en pratique, nous considérons une version tronquée de celui-ci en un seuil J . Ainsi, nous estimons $a(x)$ et b_x par la minimisation du critère suivant :

$$\min_{a,b_1,\dots,b_J \in \mathcal{H}} \sum_{i=1}^n \left\| Y_i - a - \sum_{j=1}^J c_{ij} b_j \right\|_{\mathcal{H}}^2 K(h^{-1}\|x - X_i\|_{\mathcal{F}}), \text{ où } b_j = b_x(v_j) \text{ pour } j = 1, \dots, J.$$

Les estimateurs sont donc explicitement donnés par :

$$\widetilde{A}(x) = (Q'_B K Q_B)^{-1} (Q'_B K Y), \text{ où } Q_B = \begin{pmatrix} 1 & c_{11} & \dots & c_{1J} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & c_{n1} & \dots & c_{nJ} \end{pmatrix},$$

avec $K = \text{diag}(K(h^{-1}\|x - X_1\|_{\mathcal{F}}), \dots, K(h^{-1}\|x - X_n\|_{\mathcal{F}}))$, $Y' = (Y_1, \dots, Y_n)$ et $\widetilde{A}(x)' = (\widehat{a}(x), \widehat{b}_1, \dots, \widehat{b}_J)$.

3. Erreur d'estimation

Dans le but d'établir la convergence presque complète de notre estimateur, nous évaluons l'erreur d'estimation. Cette dernière s'avère également utile pour établir la convergence en norme quadratique ou la normalité asymptotique. Pour ce faire, nous introduisons les notations suivantes, pour $k, j = 1, \dots, J$:

$$\left\{ \begin{array}{ll} S_{n,k,j} = \frac{1}{nh^2\phi_x(h)} \sum_{i=1}^n c_{ik}c_{ij}K_i(x) & \text{et } S_{n,k,0} = \frac{1}{nh\phi_x(h)} \sum_{i=1}^n c_{ik}K_i(x) \\ T_{n,j} = \frac{1}{nh\phi_x(h)} \sum_{i=1}^n c_{ij}K_i(x)Y_i & \text{et } T_{n,0} = \frac{1}{n\phi_x(h)} \sum_{i=1}^n K_i(x)Y_i \\ T_{n,j}^* = \frac{1}{nh\phi_x(h)} \sum_{i=1}^n c_{ij}K_i(x)(Y_i - r(X_i)) & \text{et } T_{n,0}^* = \frac{1}{n\phi_x(h)} \sum_{i=1}^n K_i(x)(Y_i - r(X_i)) \\ e_{n,j} = \frac{1}{nh\phi_x(h)} \sum_{i=1}^n c_{ij}K_i(x)\rho_x(X_i - x, X_i - x) & \text{et } e_{n,0} = \frac{1}{n\phi_x(h)} \sum_{i=1}^n K_i(x)\rho_x(X_i - x, X_i - x) \\ e_{n,j}^* = S_{n,j,0} = \frac{1}{nh\phi_x(h)} \sum_{i=1}^n c_{ij}K_i(x) & \text{et } e_{n,0}^* = S_{n,0,0} = \frac{1}{n\phi_x(h)} \sum_{i=1}^n K_i(x) \end{array} \right.$$

où $\phi_x(h) = P(X \in B(x, h))$ est supposée strictement positive et $K_i(x) = K(h^{-1}\|x - X_i\|_{\mathcal{F}})$. Ainsi, nous pouvons écrire :

$$\widetilde{A}(x)_h = (S_n)^{-1} (T_n), \text{ où } S_n = (S_{n,k,j})_{k,j=0,\dots,J}, T_n = (T_{n,j})_{j=0,\dots,J} \text{ et } \widetilde{A}(x)'_h = (\widehat{a}(x), h\widehat{b}_1, \dots, h\widehat{b}_J).$$

En vertu de l'approximation (2) et du fait que l'erreur de la projection sur une base orthonormée d'un espace de Hilbert est de l'ordre de $O(J^{-1})$, nous avons :

$$r(X_i) = a(x) + \sum_{j=1}^J c_{ij}b_x(v_j) + \rho_x(X_i - x, X_i - x) + O(J^{-1}).$$

Par ailleurs, considérons $T_n^* = (T_{n,j}^*)_{j=0,\dots,J}$, $e_n = (e_{n,j})_{j=0,\dots,J}$ et $e_n^* = (e_{n,j}^*)_{j=0,\dots,J}$. On peut écrire :

$$T_n^* = T_n - (T_n - T_n^*) = S_n \widetilde{A}(x)_h - S_n A(x) + e_n + O(J^{-1})e_n^*,$$

où $A'(x) = (a(x), hb_1, \dots, hb_J)$. Par conséquent : $\widetilde{A}(x)_h - A(x) = S_n^{-1} T_n^* - S_n^{-1} e_n - O(J^{-1})S_n^{-1} e_n^*$.

Nous déduisons l'erreur d'estimation suivante :

$$\widehat{a}(x) - a(x) = e'_1 \left(S_n^{-1} T_n^* - S_n^{-1} e_n - O(J^{-1})S_n^{-1} e_n^* \right),$$

où e'_1 désigne le vecteur transposé du premier vecteur de la base canonique de \mathbb{R}^J .

4. La convergence uniforme presque complète

Afin d'énoncer notre résultat principal, nous aurons besoin des hypothèses suivantes :

- (H0) $S_{\mathcal{F}} \subset \bigcup_{k=1}^{d_n} B(x_k, r_n)$ où $x_k \in \mathcal{F}$ et r_n (resp. d_n) est une suite de nombres réels positifs ;
- (H1) il existe une fonction dérivable $\phi(\cdot)$ vérifiant : $0 < C\phi(h) \leq \phi_x(h) \leq C'\phi(h) < \infty$ et $\exists \eta_0 > 0, \forall \eta < \eta_0, \phi'(\eta) < C$ où C et C' sont des constantes strictement positives et ϕ' est la dérivée de ϕ ;
- (H2) il existe $m \geq 2$ et $C > 0$ tels que $E(\|Y\|^m | X = x) < C$ et $E(\|r(X)\|^m | X = x) < C$;
- (H3) pour tout $x \in S_{\mathcal{F}}$, l'opérateur ρ_x est continu et vérifie $\sup_{x \in S_{\mathcal{F}}} \|\rho_x\| < C$;
- (H4) le noyau K est une fonction positive, différentiable, à support compact inclus dans $[0, 1]$ et il existe une constante $C > 0$ telle que, pour tout x, y : $|K(x) - K(y)| \leq C\|x - y\|$;

(H5) pour $r_n = O(\log n/n)$, la suite d_n vérifie :

$$\frac{(\log n)^2}{n\phi(h)} < \log(d_n) < \frac{n\phi(h)}{\log n} \text{ et } \exists \beta > 1, \sum_{n=1}^{\infty} \exp \left\{ (1-\beta)\psi_{S_{\mathcal{F}}} \left(\frac{\log n}{n} \right) \right\} < \infty,$$

où $\psi_{S_{\mathcal{F}}}(\cdot)$ désigne la ε -entropie de Kolmogorov de $S_{\mathcal{F}}$.

Théorème 1. Sous les hypothèses (H0)–(H5), on a :

$$\sup_{x \in S_{\mathcal{F}}} \|\widehat{a}(x) - a(x)\|_{\mathcal{H}} = O(J^{-1}) + O(h^2) + O_{p.co.} \left(\sqrt{\frac{\log d_n}{n\phi(h)}} \right).$$

Schéma de la démonstration : il suffit d'étudier la convergence presque complète de chaque composante dans la matrice S_n et dans les vecteurs T_n^* et e_n . La démonstration détaillée peut être obtenue sur simple demande.

5. Quelques remarques conclusives

Notons que la vitesse de convergence de notre estimateur dépend de paramètres liés à la structure topologique de l'espace fonctionnel des variables et/ou à la méthode de construction de notre estimateur. Nos résultats confirment la supériorité de la méthode locale linéaire sur la méthode à noyau. En effet, la méthode à noyau fournit un biais d'ordre $O(h)$; en revanche, avec la méthode locale linéaire, on obtient un biais d'ordre $O(h^2)$. Par ailleurs, sur la partie variance, on a le même constat que dans le cas multidimensionnel. Seulement, dans notre cadre d'étude, le terme de variance est exprimé en fonction de la concentration $\phi(h)$ et de l'entropie d_n , qui dépend, à son tour, de $S_{\mathcal{F}}$ et de la suite r_n . Ces deux fonctions sont liées à la structure topologique de la variable explicative. Le terme additionnel $O(J^{-1})$, quant à lui, est lié à la méthode de construction de l'estimateur et contrôle l'erreur de projection orthogonale sur le sous-espace engendré par v_1, \dots, v_J . Ce terme additionnel n'est point un facteur de détérioration, car un simple choix de J , i.e. $J = O(\sqrt{n/\log d_n})$, permet d'avoir une expression standard de la vitesse de convergence : $O\left(h^2 + \sqrt{\log d_n/n\phi(h)}\right)$.

Cependant, dans cette note, nous avons opté pour une présentation permettant d'exploiter ce terme à travers l'expression de la vitesse de convergence. Ceci met en évidence l'importance du choix de J . Un choix judicieux de ce paramètre sera basé sur une balance entre la précision et la rapidité dans le calcul de l'estimateur. En effet, un choix trop petit de J donne un estimateur rapide à calculer, mais qui a une vitesse de convergence lente, et vice versa. Dans le même ordre d'idée, le meilleur choix du paramètre de lissage h est obtenu en minimisant la partie dominante de la vitesse de convergence $O\left(h^2 + \sqrt{\log d_n/n\phi(h)}\right)$. Mais, cette quantité n'étant pas explicitement connue, nous proposons donc d'utiliser le critère classique de validation croisée défini par : $\frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{a}_i(X_i))$, où \widehat{a}_i est l'estimateur validé croisé, obtenu sans l'observation X_i . L'optimalité asymptotique de ce critère est une question ouverte, que nous étudierons ultérieurement.

Références

- [1] A. Baillo, A. Grané, Local linear regression for functional predictor and scalar response, *J. Multivar. Anal.* 100 (2009) 102–111.
- [2] J. Barrientos-Marin, F. Ferraty, P. Vieu, Locally modelled regression and functional data, *J. Nonparametr. Stat.* 22 (2010) 617–632.
- [3] A. Berlinet, A. Elamine, A. Mas, Local linear regression for functional data, *Ann. Inst. Stat. Math.* 63 (2011) 1047–1075.
- [4] E. Boj, P. Delicado, J. Fortiana, Distance-based local linear regression for functional predictors, *Comput. Stat. Data Anal.* 54 (2010) 429–437.
- [5] D. Bosq, *Linear Processes in Function Spaces: Theory and Applications*, Lecture Notes in Statistics, vol. 149, Springer, 2000.
- [6] A. Cuevas, A partial overview of the theory of statistics with functional data, *J. Stat. Plan. Inference* 147 (2014) 1–23.
- [7] S. Dabo-Niang, N. Rhomari, Kernel regression estimation in a Banach space, *J. Stat. Plan. Inference* 139 (2009) 1421–1434.
- [8] J. Demongeot, A. Laksaci, F. Madani, M. Rachdi, Functional data: local linear estimation of the conditional density and its application, *Statistics* 47 (2013) 26–44.
- [9] J. Fan, I. Gijbels, *Local Polynomial Modelling and Its Applications*, Chapman & Hall, London, 1996.
- [10] F. Ferraty, A. Laksaci, A. Tadj, P. Vieu, Kernel regression with functional response, *Electron. J. Stat.* 5 (2011) 159–171.
- [11] F. Ferraty, A. Laksaci, A. Tadj, P. Vieu, Estimation de la fonction de régression pour variable explicative et réponses fonctionnelles dépendantes, *C. R. Acad. Sci. Paris, Ser. I* 350 (2012) 717–720.
- [12] F. Ferraty, I. Van Keilegom, P. Vieu, Regression when both response and predictor are functions, *J. Multivar. Anal.* 109 (2012) 10–28.
- [13] F. Ferraty, P. Vieu, *Nonparametric Functional Data Analysis. Theory and Practice*, Springer Series in Statistics, Springer-Verlag, New York, 2006.
- [14] A. Goia, P. Vieu, An introduction to recent advances in high/infinite dimensional statistics, *J. Multivar. Anal.* 146 (2016) 1–6, <http://dx.doi.org/10.1016/j.jmva.2015.12.001>.
- [15] T. Hsing, R. Eubank, *Theoretical Foundations of Functional Data Analysis, With an Introduction to Linear Operators*, Wiley Series in Probability and Statistics, John Wiley & Sons, Chichester, UK, 2015.
- [16] J.O. Ramsay, B.W. Silverman, *Applied Functional Data Analysis. Methods and Case Studies*, Springer Series in Statistics, Springer-Verlag, New York, 2002.
- [17] J. Zhang, *Analysis of Variance for Functional Data*, Monographs on Statistics and Applied Probability, vol. 127, CRC Press, Boca Raton, FL, USA, 2014.