Available online at www.sciencedirect.com
ScienceDirect
C. R. Acad. Sci. Paris, Ser. I 346 (2008) 339–342
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistique

Convergence de l'estimateur à noyau des k plus proches voisins en régression fonctionnelle non-paramétrique

Florent Burba, Frédéric Ferraty, Philippe Vieu

I.M.T., UMR 5219, Université Paul-Sabatier, 31062 Toulouse, France

Reçu le 26 octobre 2007 ; accepté après révision le 29 janvier 2008

Disponible sur Internet le 20 février 2008

Présenté par Paul Deheuvels

Résumé

Cet article s'intéresse à la méthode des k plus proches voisins lorsqu'on régresse une variable réelle sur une variable fonctionnelle (i.e. à valeurs dans un espace de dimension infinie). Plus précisément, on considère un estimateur à noyau de la régression construit à partir d'une fenêtre locale permettant de prendre en compte exactement les k plus proches voisins. Bien que couramment utilisée en pratique pour le traitement de données fonctionnelles, cette méthode n'a aucun résultat théorique à ce jour. Cette Note a donc pour objectif de montrer la convergence presque-complète ponctuelle de cet estimateur non-paramétrique fonctionnel. ***Pour citer cet article : F. Burba et al., C. R. Acad. Sci. Paris, Ser. I 346 (2008).***

Abstract

Convergence of k nearest neighbor kernel estimator in nonparametric functional regression. This note focuses on the k nearest neighbor method when one regresses a real random variable on a functional random variable (i.e. valued in an infinite-dimensional space). More precisely, we consider a kernel estimator of the regression based on a local bandwidth using exactly the k nearest neighbors. Although it is frequently used in functional data analysis, this method has not given any theoretical result so far. The aim of this Note is to show the pointwise almost-complete convergence of the k nearest neighbor kernel estimator in nonparametric functional regression. ***To cite this article: F. Burba et al., C. R. Acad. Sci. Paris, Ser. I 346 (2008).***

1. Introduction

Le cadre de ce travail est la régression non-paramétrique d'une variable aléatoire (v.a.) réelle Y sur une v.a. fonctionnelle (v.a.f.) $\mathcal{X}$ (v.a. à valeurs dans un espace de dimension infinie). On s'intéresse plus précisément à l'estimation de la fonction de régression via la méthode des k plus proches voisins (kNN). En dimension finie (i.e. $\mathcal{X}$ est à valeurs dans $\mathbb{R}^p$), l'estimateur kNN à noyau de la régression a été largement étudié (voir Breiman et al. [3], Collomb, [6], Györfi, [10], Mack, [12], ou encore Bhattacharya et Mack, [1]). Dans le cadre fonctionnel, cette méthode présente de nombreux avantages, principalement celui de respecter la structure locale des données, ce qui est primordial en dimension infinie. Elle est couramment utilisée en pratique (voir Ferraty et Vieu, [9]) et s'avère simple à manipuler

Adresses e-mail : burba@cict.fr (F. Burba), ferraty@cict.fr (F. Ferraty), vieu@cict.fr (P. Vieu).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2008.01.022

car l'utilisateur n'a qu'un seul paramètre à contrôler (le nombre k de plus proches voisins), ce paramètre k prenant ses valeurs dans un ensemble fini. De plus, elle permet de construire en tout point un voisinage adapté aux données. Cependant, cette méthode fait apparaître une première difficulté : la fenêtre locale des k plus proches voisins est aléatoire. Une seconde difficulté, liée à la nature fonctionnelle des données, est que l'on s'affranchit de l'existence d'une densité pour la v.a.f. $\mathcal{X}$. L'objectif de cette Note est d'apporter une première justification théorique à l'utilisation courante de la méthode kNN en régression fonctionnelle non-paramétrique.

2. Modèle et estimateur

2.1. Modèle

Soient $(\mathcal{X}_i, Y_i)_{i=1,\dots,n}$ n paires indépendantes et identiquement distribuées comme $(\mathcal{X}, Y)$ et à valeurs dans $E \times \mathbb{R}$ où (E, d) est un espace semi-métrique et non nécessairement de dimension finie. On s'intéresse au modèle de régression fonctionnelle non-paramétrique suivant :

$$Y = r(\mathcal{X}) + \epsilon \quad \text{avec } \mathbb{E}[\epsilon|\mathcal{X}] = 0.$$

L'objectif est donc ici d'estimer l'opérateur $r(\cdot) = \mathbb{E}[Y|\mathcal{X} = \cdot]$.

2.2. Estimateur

Soit $\chi \in E$ fixé. La méthode des k plus proches voisins consiste à considérer les k variables $\mathcal{X}_i$ les plus proches de χ au sens de la semi-métrique d. Pour k fixé, on définit ainsi $H_{n,k}(\chi)$ comme étant le plus petit réel h pour lequel $\sum_{i=1}^{n} 1_{B(\chi,h)}(\mathcal{X}_i) = k$. $H_{n,k}(\chi)$ est une variable aléatoire de E dans $\mathbb{R}^+$ qui dépend de $(\mathcal{X}_1, \dots, \mathcal{X}_n)$. L'estimateur à noyau des k plus proches voisins de r en χ est défini de la façon suivante

$$\hat{r}_{kNN}(\chi) = \sum_{i=1}^{n} Y_i \omega_{i,n}(\chi) \quad \text{avec } \omega_{i,n}(\chi) = \frac{K(H_{n,k}(\chi)^{-1} d(\chi, \mathcal{X}_i))}{\sum_{i=1}^{n} K(H_{n,k}(\chi)^{-1} d(\chi, \mathcal{X}_i))}$$

où K est un noyau asymétrique. De plus, dès que le support de K est $[0, 1]$, il est clair que $\hat{r}_{kNN}(\chi) = \sum_{j=1}^{k} Y_{i_j} \omega_{i_j,n}(\chi)$ où $\{i_1, \dots, i_k\} = \{i \mid \mathcal{X}_i \in B(\chi, H_{n,k}(\chi))\}$.

3. Convergence ponctuelle de $\hat{r}_{kNN}$

On montre ici la convergence presque-complète (*pco*) ponctuelle de $\hat{r}_{kNN}(\chi)$ vers $r(\chi)$ pour $\chi \in E$ fixé. Le modèle non-paramétrique utilisé se caractérise par des contraintes sur la régularité de r :

$$r \in \mathcal{C}_E^0 = \left\{ f\colon E \to \mathbb{R} \;\middle|\; \lim_{d(\chi,\chi')\to 0} f(\chi) = f(\chi') \right\}. \tag{1}$$

On fait ensuite des hypothèses sur la loi du couple $(\mathcal{X}, Y)$ et sur l'estimateur $\hat{r}_{kNN}$:

- (H_1) Hypothèse de concentration agissant sur la v.a. fonctionnelle $\mathcal{X}$: $\forall \epsilon > 0,\ P(\mathcal{X} \in B(\chi, \epsilon)) = \varphi_\chi(\epsilon) > 0$ avec $\varphi_\chi(\cdot)$ continue au voisinage de 0 et $\varphi_\chi(0) = 0$.
- (H_2) Hypothèse sur les moments conditionnels de la v.a. réponse Y : $\forall m \geqslant 2, \mathbb{E}[|Y|^m|\mathcal{X} = \chi] = \sigma_m(\chi) < \infty$ avec $\sigma_m(\cdot)$ continue en χ.
- (H_3) Hypothèse sur le noyau K : il existe des constantes $0 < C_1 < C_2 < \infty$ telles que $C_1 1_{[0,1]} \leqslant K \leqslant C_2 1_{[0,1]}$.
- (H_4) Hypothèse sur le nombre k de voisins : $k = k_n$ est une suite de réels positifs telle que $\frac{k}{n} \to 0$ et $\frac{\log n}{k} \to 0$.

Théorème 3.1. *Dans le modèle* (1) *et si* (H_1)–(H_4) *sont vérifiées, alors* :

$$\hat{r}_{kNN}(\chi) \xrightarrow{(pco)} r(\chi)$$

La preuve intégrale de ce résultat, sous des hypothèses légèrement plus générales permettant notamment de considérer des noyaux K continus, peut être obtenue sur simple demande.

Eléments de preuve. La première difficulté vient du fait que la fenêtre $H_{n,k}(\chi)$ est aléatoire. Ainsi, nous n'avons pas au numérateur et au dénominateur de $\hat{r}_{kNN}(\chi)$ des sommes de variables indépendantes. Pour résoudre ce problème, l'idée est d'encadrer judicieusement $H_{n,k}(\chi)$ par deux fenêtres non-aléatoires. On utilise pour cela le résultat suivant établi par Collomb [6] :

Lemme 3.2. *Soit $(D_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles. Si, $\forall\beta\in\,]0,1[$ fixé, il existe deux suites de variables aléatoires réelles $(D_n^-)_{n\in\mathbb{N}}$ et $(D_n^+)_{n\in\mathbb{N}}$ telles que* :

(L_1) $D_n^- \leqslant D_n^+$ $\forall n\in\mathbb{N}$ *et* $1_{\{D_n^-\leqslant D_n\leqslant D_n^+\}} \xrightarrow{(pco)} 1$;
(L_2) $\frac{\sum_{i=1}^n K((D_n^-)^{-1}d(\chi,\mathcal{X}_i))}{\sum_{i=1}^n K((D_n^+)^{-1}d(\chi,\mathcal{X}_i))} \xrightarrow{(pco)} \beta$;
(L_3) $c_n(D_n^-) \xrightarrow{(pco)} c$ *et* $c_n(D_n^+) \xrightarrow{(pco)} c$ *où* $c_n(T) = \frac{\sum_{i=1}^n Y_i K(T^{-1}d(\chi,\mathcal{X}_i))}{\sum_{i=1}^n K(T^{-1}d(\chi,\mathcal{X}_i))}$.

Alors, $c_n(D_n) \xrightarrow{(pco)} c$.

On pose ici $D_n = H_{n,k}(\chi)$, on a $c_n(H_{n,k}(\chi)) = \hat{r}_{kNN}(\chi)$ et $c = r(\chi)$. Soit ensuite $\beta\in\,]0,1[$. On choisit D_n^+ et D_n^- des fenêtres non-aléatoires telles que $\varphi_\chi(D_n^-) = \sqrt{\beta}\frac{k}{n}$ et $\varphi_\chi(D_n^+) = \frac{1}{\sqrt{\beta}}\frac{k}{n}$, ce choix étant possible grâce à (H_1). La vérification de (L_2) et (L_3) se fait directement d'après les travaux de Ferraty et Vieu [9] sur l'estimateur à noyau à fenêtre fixe.

La deuxième difficulté vient du fait qu'aucune hypothèse de densité sur $\mathcal{X}$ n'a été faite. Ainsi, la démarche suivie par Collomb [6] pour vérifier (L_1) en dimension finie n'est plus transposable ici et l'on procède comme suit. Pour $\epsilon > 0$, on a

$$P\big(|1_{\{D_n^-\leqslant H_{n,k}(\chi)\leqslant D_n^+\}} - 1| > \epsilon\big) \leqslant P\big(H_{n,k}(\chi) < D_n^-\big) + P\big(H_{n,k}(\chi) > D_n^+\big),$$

ce qui se réécrit :

$$P\big(|1_{\{D_n^-\leqslant H_{n,k}(\chi)\leqslant D_n^+\}} - 1| > \epsilon\big) \leqslant P\left(\sum_{i=1}^n 1_{B(\chi,D_n^-)}(\mathcal{X}_i) > k\right) + P\left(\sum_{i=1}^n 1_{B(\chi,D_n^+)}(\mathcal{X}_i) < k\right).$$

En utilisant des inégalités exponentielles de type Chernoff sur les sommes de variables aléatoires de Bernoulli (voir Chernoff, [5]), on peut majorer chacune des deux probabilités de la somme précédente par un terme de type $n^{-a\frac{k}{\log n}}$ où $a > 0$, ce qui permet d'établir que $\sum_{n\in\mathbb{N}} P(|1_{\{D_n^-\leqslant H_{n,k}(\chi)\leqslant D_n^+\}} - 1| > \epsilon) < \infty$. □

4. Remarques et commentaires

Un point important de cet article est le fait de travailler directement avec les probabilités de petite boule. Ceci évite d'avoir à supposer l'existence d'une mesure dominante, chose toujours délicate dans des espaces de dimension infinie (voir Dabo-Niang, [7], pour une large discussion à ce sujet). De plus, on a supposé que E n'était pas nécessairement de dimension finie. Le résultat présenté étend ainsi les résultats déjà connus en dimension finie (voir par exemple Györfi et al., [11]) non seulement à des v.a.f. mais aussi à des v.a. multivariées n'admettant pas nécessairement de densité.

On remarque aussi que l'on peut directement utiliser la méthode kNN pour faire de la discrimination de courbes. En effet, si $Z\in\{1,\dots,G\}$ représente un numéro de groupe et si on pose $Y = 1_{\{Z=g\}}$ pour $g\in\{1,\dots,G\}$, on a $P(Z=g|\mathcal{X}) = \mathbb{E}[Y|\mathcal{X}]$. De ce point de vue, cet article peut aussi être vu comme une extension à la dimension infinie des résultats de Györfi [10] (voir aussi Devroye et Wagner, [8], pour un éventail de résultats en dimension finie sur la discrimination par méthode kNN, Biau et al., [2], et Cérou et Gudayer, [4], pour des résultats récents en dimension infinie).

Enfin, les principales perspectives de poursuite de ce travail concernent les vitesses de convergence presque-complète et la normalité asymptotique de $\hat{r}_{kNN}$ ainsi que le choix automatique du paramètre k.

Remerciements

Ces travaux s'inscrivent dans la dynamique que le groupe STAPH de Toulouse impulse autour des divers aspects de la Statistique Fonctionnelle et Opératorielle et dont tous les participants sont ici remerciés chaleureusement pour leurs commentaires permanents. Les auteurs remercient aussi les rapporteurs pour leur lecture attentive et leurs propositions concrètes d'amélioration.

Références

[1] P.K. Bhattacharya, Y.P. Mack, Weak convergence of k-NN density and regression estimators with varying k and applications, Ann. Statist. 15 (1987) 976–994.
[2] G. Biau, F. Bunea, M.H. Wegkamp, Functional classification in Hilbert spaces, IEEE Trans. Inform. Theory 51 (2005) 2163–2172.
[3] L. Breiman, W. Meisel, E. Purcell, Variable kernel estimates of multivariate densities, Technometrics 19 (1977) 135–144.
[4] F. Cérou, A. Gudayer, Nearest neighbor classification in infinite dimension, ESAIM P&S 10 (2006) 340–355.
[5] H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, Ann. Math. Stat. 23 (1952) 493–507.
[6] G. Collomb, Estimation de la régression par la méthode des k points les plus proches avec noyau : quelques propriétés de convergence ponctuelle, in : Nonparametric Asymptotic Statistics, in : Lecture Notes in Mathematics, vol. 821, Springer-Verlag, 1980, pp. 159–175.
[7] S. Dabo-Niang, Sur l'estimation de la densité en dimension infinie : applications aux diffusions, Thèse de Doctorat, Université Paris 6, 2002.
[8] L. Devroye, T.J. Wagner, Nearest neighbor methods in discrimination, in: Classification, Pattern Recognition and Reduction of Dimensionnality, in: Handbook of Statistics, vol. 2, North-Holland, Amsterdam, 1982, pp. 193–197.
[9] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice, Springer Series in Statistics, Springer, New York, 2006.
[10] L. Györfi, The rate of convergence of kNN-regression estimation and classification, IEEE Trans. Inform. Theory 27 (1981) 500–509.
[11] L. Györfi, M. Kohler, A. Krzyzak, H. Walk, A Distribution-Free Theory of Nonparametric Regression, Springer Series in Statistics, Springer-Verlag, New York, 2002.
[12] Y.P. Mack, Local properties of kNN-regression estimates, J. Alg. Disc. Meth. 2 (1981) 311–323.