C. R. Acad. Sci. Paris, Ser. I 356 (2018) 1036–1039

Contents lists available at ScienceDirect

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

www.sciencedirect.com

Statistics

*k*NN local linear estimation of the conditional cumulative distribution function: Dependent functional data case

Analyse des séries temporelles fonctionnelles : estimation locale linéaire et par les k plus proches voisins de la fonction de répartition conditionnelle

Ibrahim M. Almanjahie [a], Zouaoui Chikr Elmezouar [a], Ali Laksaci [a], Mustapha Rachdi [b]

[a] *Department of Mathematics, College of Science, King Khalid University, Abha, 61413, Saudi Arabia*
[b] *Université Grenoble Alpes, Laboratoire AGEIS, EA 7407, AGIM–TIMB Team, UFR SHS, BP 47, 38040 Grenoble cedex 09, France*

ARTICLE INFO

Article history:
Received 11 June 2018
Accepted after revision 10 September 2018
Available online 27 September 2018

Presented by Paul Deheuvels

ABSTRACT

Let $(X_i, Y_i)_{i=1,\dots,n}$ be a sequence of strongly mixing random vectors valued in $\mathcal{F}\times\mathbb{R}$, where $\mathcal{F}$ is a Hilbert space. This note deals with the problem of the local linear estimation of the conditional cumulative distribution function of Y_i given X_i. Then, the main goals of this note are (i) to construct a fast *k*NN local linear estimate of the conditional distribution function, and (ii) to prove its strong consistency in the functional time series analysis situation.

RÉSUMÉ

Soit $(X_i, Y_i)_{i=1,\dots,n}$ une suite de vecteurs aléatoires, fortement mélangeants, à valeurs dans $\mathcal{F}\times\mathbb{R}$, où $\mathcal{F}$ est un espace de Hilbert. Dans cette note, nous nous intéressons au problème d'estimation locale linéaire de la distribution conditionnelle de Y_i sachant X_i. Les objectifs principaux de cette note sont donc (i) de construire un estimateur local linéaire rapide, en introduisant la technique des k plus proches voisins, de la distribution conditionnelle et (ii) de prouver sa consistance forte dans le cadre des séries temporelles conditionnelles.

1. Introduction

In the last decades, Functional Data Analysis (FDA) has known continuous developments. It is among the most active fields of research in statistic (see for instance [1], [7] and [12] for some recent and general studies). The nonparametric functional data analysis was started in the beginning of this century. It was popularized by the monograph [10]. Since the

E-mail addresses: imalanjhi@kku.edu.sa (I.M. Almanjahie), chikertime@yahoo.fr (Z. Chikr Elmezouar), alilak@yahoo.fr (A. Laksaci), mustapha.rachdi@univ-grenoble-alpes.fr (M. Rachdi).

https://doi.org/10.1016/j.crma.2018.09.001
1631-073X/

publication of this monograph, various nonparametric models have been developed in FDA (see, for instance, [11] and [18] for a state of the art and a deeper discussion on this topic). Notice that among the widely developed approaches, in the last few years, we find the functional local linear method (LLM). The principal motivation of this subject is the superiority of this approach over the classical kernel method (CKM). In particular, the LLM has a small bias compared to the kernel method (see [9], for the finite dimension framework, and [2], for the FDA setup). In fact, Ba'ıllo and Grané [2] constructed a local linear estimate (LLE) of the regression operator when the explanatory variable belongs to an Hilbert space, whereas Barrientos-Marin, Ferraty, and Vieu's LLE [3] is defined on a Banach space. On the other hand, Berlinet, Elamine, and Mas's version [4] is based on the use of the inverse of the local covariance operator of the functional regressor, whereas Zhou and Lin [20] studied the asymptotic normality of the LLE of the regression operator.

Furthermore, the conditional distribution function (CDF) was studied in [15], which established the almost complete (a.co.) consistency of a LLE of this CDF, and in which the asymptotic result is obtained under a spatial dependency condition. We return to [8] for the mean quadratic consistency of the LLE of the CDF in FDA. All these cited works used the so-called kernel local linear method; however, in this note, we use a new approach that is based on a mixing of the kNN method and LLM ideas to estimate the CDF.

Recall that the kNN method has more practical advantages than the CKM. In particular, it allows us to construct an attractive estimate that is adapted to the local structure of the data (see [6] for more motivations on this topic). The kNN method has been studied in the nonparametric FDA setting by many authors (see, for instance, [5], [16] and [17] for previous works on the functional kNN method and [13] and [14] for recent advances and references on FDA).

Whilst the most studies on the functional kNN method are oriented toward the regression estimation by a local constant method, we consider, here, a more efficient estimate of the CDF by the LLM. Precisely, in order to benefit from the nice features of both the LLM and the kNN procedures, we combine the two approaches. In other words, we construct a new estimate of the CDF and we study its asymptotic properties. Hence, our main asymptotic result is the establishment of the a.co. consistency (with rate) of the constructed estimate. This asymptotic result is obtained under an α-mixing condition.

It must be noticed that the smoothing parameter, in the kNN method, is a random variable that makes the asymptotic study of this method more difficult than the local constant method, where the smoothing parameter is a scalar. This difficulty becomes more complicated in the functional time series setting. Nevertheless, the functional time series case is a more realistic situation than the independent one. Typically, the functional time series data can be constructed from a continuous-time stochastic process. For instance, our approach can be applied to predict future values of some continuous-time process by cutting the whole past of this process into continuous paths.

This paper is organized as follows. Section 2 is devoted to the presentation of the LLE estimate of the CDF. Then, the main result is given in Section 3.

2. The model and its estimation

Let $(X_i, Y_i)_{i=1,\dots,n}$ be a sequence of strongly α-mixing and identically distributed random vectors that are valued in $\mathcal{F} \times \mathbb{R}$, where $\mathcal{F}$ is a Hilbert space.

In what follows, we fix a curve $x \in \mathcal{F}$ and we consider a given neighborhood $\mathcal{N}_x$ of x. We assume that the regular version of the CDF $F(\cdot, x)$ of Y given $X = x$ exists. Moreover, we assume that $F(\cdot, x)$ has a continuous density $f(\cdot, x)$ with respect to Lebesgue's measure over $\mathbb{R}$.

In FDA, there are several ways for extending the LLM ideas (see, for instance, [2], [3] and [8] and references therein). But, all local linear fitting requires a smoothing condition that allows us to approximate locally the nonparametric model by a linear function. Then, we estimate the function $F(\cdot, x)$ as follows. We assume that, for a fixed $(y, x) \in \mathbb{R} \times \mathcal{F}$, the function $F(y, x)$ is smoothed enough to be locally approximated by a linear function. That is, for all $x_0 \in \mathcal{N}_x$:

$$F(y, x_0) = a_{yx} + b_{yx}(x_0 - x) + \rho_{yx}(x_0 - x, x_0 - x) + o(\|x_0 - x\|^2), \tag{1}$$

where b_{yx} (respectively, ρ_{yx}) is a linear (respectively, a bilinear continuous) operator from $\mathcal{F}$ into $\mathbb{R}$ (respectively, $\mathcal{F} \times \mathcal{F}$ into $\mathbb{R}$).

In order to construct the LLE of the CDF $F(y, x)$, we use the fact that:

$$\mathbb{P}\left(Y \le y \middle| X = x\right) = \mathbb{E}\left[\mathbb{1}_{\{Y \le y\}} \middle| X = x\right],$$

where $\mathbb{1}_A$ denotes the indicator function on the set A. So, the kNN estimates of the operators a_{yx} and b_{yx}, in the formula (1), are the minimizers of the following rule:

$$\min_{a,b} \sum_{i=1}^{n} \left(\mathbb{1}_{\{Y_i \le y\}} - a - b(X_i - x)\right)^2 K\left(\frac{\|x - X_i\|}{h_k}\right),$$

where K is a kernel function and $h_k = \min\{h \in \mathbb{R}^+ \text{ such that } \sum_{i=1}^{n} \mathbb{1}_{B(x,h)}(X_i) = k\}$.

Then, by using the same ideas as in [2], we built an estimate of the CDF $F(y,x)$ as follows. Let

$$Q_B = \begin{pmatrix} 1 & c_{11} & \dots & c_{1J} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & c_{n1} & \dots & c_{nJ} \end{pmatrix},$$

where c_{ij} are the coefficients of $(X_i - x)$ in a truncated version of the basis $(v_j)_{1\le j\le J}$ of $\mathcal{F}$ and put

$$K = \mathrm{diag}(K(h_k^{-1}\|x - X_1\|), \dots, K(h_k^{-1}\|x - X_n\|)) \text{ and } Y' = \left(\mathbb{1}_{\{Y_1\le y\}}, \dots, \mathbb{1}_{\{Y_n\le y\}}\right).$$

Then, we assume that $(Q_B' K Q_B)$ is a nonsingular matrix to express the LLE of the CDF $F(y,x)$ by:

$$\widehat{F}(y,x) = \widehat{a}_{yx} = e_1'(Q_B' K Q_B)^{-1}(Q_B' K Y),$$

where e_1' denotes the transpose of the first vector of the canonical basis of $\mathbb{R}^J$.

3. Asymptotic properties of the estimate $\widehat{F}(y,x)$

In what follows, we denote by (y,x) a fixed point in $\mathbb{R}\times\mathcal{F}$ and $\mathcal{N}_y$ a fixed neighborhood of y. Furthermore, we assume that our nonparametric model satisfies the following conditions:

$$\text{There exists } a>2, \text{ such that } \sum n^a \alpha(n) < \infty. \tag{2}$$

$$\text{For any } r>0, \ \ \phi_x(r) := \mathbb{P}(X\in B(x,r)) > 0 \text{ is an invertible function,} \tag{3}$$

and, for all $i\neq j$:

$$0 < \sup_{i\neq j} \mathbb{P}\left[(X_i, X_j)\in B(x,h)\times B(x,h)\right] \le C(\phi_x(h))^{a/(a-1)}, \tag{4}$$

where $B(x,r) := \{z\in\mathcal{F} \text{ such that } d(z,x) < r\}$ denotes the ball centered at x and of radius r.

$$\text{There exists } 0<c<1<c'<\infty \text{ such that } \lim_{r\to 0}\frac{\phi_x(rc)}{\phi_x(r)} < 1 < \lim_{r\to 0}\frac{\phi_x(rc')}{\phi_x(r)}. \tag{5}$$

The kernel K is a differentiable function supported within $[0,1]$ for which its first derivative K' exists and such that there exist two constants C and C' such that:

$$-\infty < C' < K'(t) < C < 0 \ \text{ for } 0\le t\le 1. \tag{6}$$

The number k of neighbors is such that:

$$k \ge C n^{\frac{4}{a+1}+\eta} \ \text{ with } \ \eta>0 \ \text{ for } a>4. \tag{7}$$

Remark 3.1. It is clear that all these conditions are common in the FDA context. In particular, the dependency setting is explored through assumption (2), for the global dependency, and assumption (4), for the local dependency. Such conditions combined with assumption (7) allow us to establish the same convergence rate as in the independent and identically distributed (i.i.d.) case. Notice also that we can obtain convergence results without these assumptions; however, the convergence rate expression will depend on the covariance between the observations.

Theorem 1. *Under assumptions (2)–(7), we have:*

$$|\widehat{F}(y,x) - F(y,x)| = O(J^{-1}) + O\left(\phi_x^{-1}\left(\frac{k}{n}\right)^2\right) + O_{a.co.}\left(\sqrt{\frac{\ln n}{k}}\right), \ \textit{as} \ \min(n,J)\to\infty.$$

Remark 3.2. It appears clear that this approach improves the bias term of the local constant method. Indeed, the bias term, here, is of order $O\left(\phi_x^{-1}(k/n)^2\right)$. But, under the same regularity condition the kernel method gives a bias term of order $O\left(\phi_x^{-1}(k/n)\right)$. Thus, we deduce that the functional local linear approach keeps its advantages as in the multivariate case that is the reduction of the bias term.

Sketch of Theorem 1's proof. The basic ideas of this proof are: (i) we use the matrix formula of $\widehat{F}(y,x)$ to explicit the error $\widehat{F}(y,x) - F(y,x)$, and (ii) we combine ideas used in [2] with those used in [6] to evaluate the obtained error. Furthermore, the dependency structure is managed by the Fuk–Nagaev inequality and the covariance term is determinated by the usual techniques in [19]. In order to save space, the proof is omitted here, but it may be obtained on a simple request to one of the authors. □

Acknowledgements

The authors would like to thank the anonymous reviewer for his/her valuable comments and suggestions that improved substantially the quality of this paper. The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through research groups program under grant number R.G.P.1/50/39.

References

[1] G. Aneiros, E.G. Bongiorno, R. Cao, P. Vieu, Functional Statistics and Related Fields, Contributions to Statistics, Springer International Publishing, 2017.
[2] A. Baìllo, A. Grané, Local linear regression for functional predictor and scalar response, J. Multivar. Anal. 100 (2009) 102–111.
[3] J. Barrientos-Marin, F. Ferraty, P. Vieu, Locally modelled regression and functional data, J. Nonparametr. Stat. 22 (5) (2010) 617–632.
[4] A. Berlinet, A. Elamine, A. Mas, Local linear regression for functional data, Ann. Inst. Stat. Math. 63 (2011) 1047–1075.
[5] F. Burba, F. Ferraty, P. Vieu, *k*-nearest neighbor method in functional non-parametric regression, J. Nonparametr. Stat. 21 (2009) 453–469.
[6] G. Collomb, Estimation non paramétrique de la régression : Revue bibliographique, Int. Stat. Rev. 49 (1981) 75–93.
[7] A. Cuevas, A partial overview of the theory of statistics with functional data, J. Stat. Plan. Inference 147 (2014) 1–23.
[8] J. Demongeot, A. Laksaci, M. Rachdi, S. Rahmani, On the local linear modelization of the conditional distribution for functional data, Sankhya, Ser. A 76 (2014) 328–355.
[9] J. Fan, I. Gijbels, Local Polynomial Modelling and Its Applications, Chapman & Hall, London, 1996.
[10] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis. Theory and Practice, Springer Series in Statistics, Springer, New York, 2006.
[11] G. Geenens, Curse of dimensionality and related issues in nonparametric functional regression, Stat. Surv. 5 (2011) 30–43.
[12] A. Goia, P. Vieu, An introduction to recent advances in high/infinite dimensional statistics, J. Multivar. Anal. 146 (2016) 1–6.
[13] L. Kara-Zaitri, A. Laksaci, M. Rachdi, P. Vieu, Data-driven *k*NN estimation for various problems involving functional data, J. Multivar. Anal. 153 (2017) 176–188.
[14] N. Kudraszow, P. Vieu, Uniform consistency of *k*NN regressors for functional variables, Stat. Probab. Lett. 83 (2013) 1863–1870.
[15] A. Laksaci, M. Rachdi, S. Rahmani, Spatial modelization: local linear estimation of the conditional distribution for functional data, Spatial Statist. 6 (2013) 1–23.
[16] T. Laloë, A *k*-nearest neighbor approach for functional regression, Stat. Probab. Lett. 78 (2008) 1189–1193.
[17] H. Lian, Convergence of functional *k*-nearest neighbor regression estimate with functional responses, Electron. J. Stat. 5 (2011) 31–40.
[18] N. Ling, P. Vieu, Nonparametric modelling for functional data: selected survey and tracks for future, Statistics 52 (2018) 934–949.
[19] E. Masry, Recursive probability density estimation for weakly dependent stationary processes, IEEE Trans. Inf. Theory 32 (1986) 254–267.
[20] Z. Zhou, Z. Lin, Asymptotic normality of locally modelled regression estimator for functional data, J. Nonparametr. Stat. 28 (2016) 116–131.